Chapter 3

Describing Relationships

case study

How Faithful Is Old Faithful?

The Starnes family visited Yellowstone National Park in hopes of seeing the Old Faithful geyser erupt. They had only about four hours to spend in the park. When they pulled into the parking lot near Old Faithful, a large crowd of people was headed back to their cars from the geyser. Old Faithful had just finished erupting. How long would the Starnes family have to wait until the next eruption?

Let's look at some data. Figure 3.1 shows a histogram of times (in minutes) between consecutive eruptions of Old Faithful in the month before the Starnes family's visit. The shortest interval was 47 minutes, and the longest was 113 minutes. That's a lot of variability! The distribution has two clear peaks—one at about 60 minutes and the other at about 90 minutes.

If the Starnes family hopes for a 60-minute gap between eruptions, but the actual interval is closer to 90 minutes, the kids will get impatient. If they plan for a 90-minute interval and go somewhere else in the park, they won't get back in time to see the next eruption if the gap is only about 60 minutes. What should the Starnes family do?

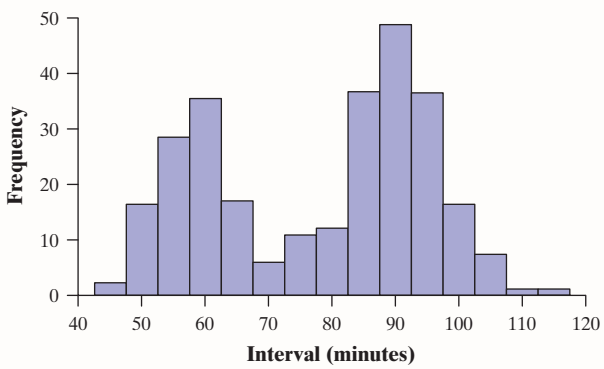


FIGURE 3.1 Histogram of the interval (in minutes) between eruptions of the Old Faithful geyser in the month prior to the Starnes family's visit.

Later in the chapter, you'll answer this question. For now, keep this in mind: to understand one variable (like eruption interval), you often have to look at how it is related to other variables.

Introduction

Investigating relationships between variables is central to what we do in statistics. When we understand the relationship between two variables, we can use the value of one variable to help us make predictions about the other variable. In Section 1.1, we explored relationships between categorical variables, such as the gender of a young person and his or her opinion about future income. The association between these two variables suggests that males are generally more optimistic about their future income than females.

In this chapter, we investigate relationships between two quantitative variables. Does knowing the number of points a football team scores per game tell us anything about how many wins it will have? What can we learn about the price of a used car from the number of miles it has been driven? Are there any variables that might help the Starnes family predict how long it will be until the next eruption of Old Faithful?

ACTIVITY CSI Stats: The case of the missing cookies

MATERIALS:

Meterstick, handprint, and math department roster (from *Teacher's Resource Materials*) for each group of three to four students; one sheet of graph paper per student

Mrs. Hagen keeps a large jar full of cookies on her desk for her students. Over the past few days, a few cookies have disappeared. The only people with access to Mrs. Hagen's desk are the other math teachers at her school. She asks her colleagues whether they have been making withdrawals from the cookie jar. No one confesses to the crime.

But the next day, Mrs. Hagen catches a break—she finds a clear handprint on the cookie jar. The careless culprit has left behind crucial evidence! At this point, Mrs. Hagen calls in the CSI Stats team (your class) to help her identify the prime suspect in "The Case of the Missing Cookies."

1. Measure the height and hand span of each member of your group to the nearest centimeter (cm). (Hand span is the maximum distance from the tip of the thumb to the tip of the pinkie finger on a person's fully stretched-out hand.)

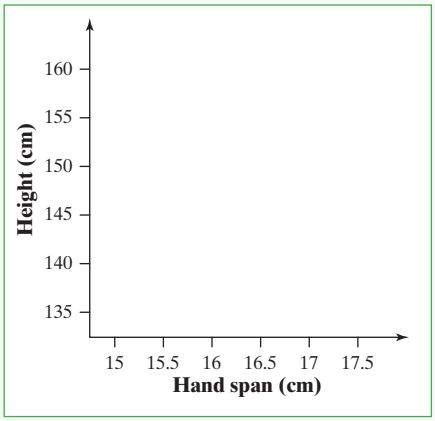
2. Your teacher will make a data table on the board with two columns, labeled as follows:

Hand span (cm)	Height (cm)

Send a representative to record the data for each member of your group in the table.

3. Copy the data table onto your graph paper very near the left margin of the page. Next, you will make a graph of these data. Begin by constructing a set of coordinate axes. Allow plenty of space on the page for your graph. Label the horizontal axis "Hand span (cm)" and the vertical axis "Height (cm)."

4. Since neither hand span nor height can be close to 0 cm, we want to start our horizontal and vertical scales at larger numbers. Scale the horizontal axis in 0.5-cm increments starting with 15 cm. Scale the vertical axis in 5-cm



increments starting with 135 cm. Refer to the sketch in the margin for comparison.

5. Plot each point from your class data table as accurately as you can on the graph. Compare your graph with those of your group members.

6. As a group, discuss what the graph tells you about the relationship between hand span and height. Summarize your observations in a sentence or two.

7. Ask your teacher for a copy of the handprint found at the scene and the math department roster. Which math teacher does your group believe is the "prime suspect"? Justify your answer with appropriate statistical evidence.

3.1 Scatterplots and Correlation

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- Identify explanatory and response variables in situations where one variable helps to explain or influences the other.
- Make a scatterplot to display the relationship between two quantitative variables.
- Describe the direction, form, and strength of a relationship displayed in a scatterplot and identify outliers in a scatterplot.
- Interpret the correlation.
- Understand the basic properties of correlation, including how the correlation is influenced by outliers.
- Use technology to calculate correlation.
- Explain why association does not imply causation.

Most statistical studies examine data on more than one variable. Fortunately, analysis of several-variable data builds on the tools we used to examine individual variables. The principles that guide our work also remain the same:

- Plot the data, then add numerical summaries.
- Look for overall patterns and departures from those patterns.
- When there's a regular overall pattern, use a simplified model to describe it.

Explanatory and Response Variables

We think that car weight helps explain accident deaths and that smoking influences life expectancy. In these relationships, the two variables play different roles. Accident death rate and life expectancy are the **response variables** of interest. Car weight and number of cigarettes smoked are the **explanatory variables**.

DEFINITION: Response variable, explanatory variable

A **response variable** measures an outcome of a study. An **explanatory variable** may help explain or predict changes in a response variable.

You will often see explanatory variables called *independent variables* and response variables called *dependent variables.* Because the words "independent" and "dependent" have other meanings in statistics, we won't use them here.

It is easiest to identify explanatory and response variables when we actually specify values of one variable to see how it affects another variable. For instance, to study the effect of alcohol on body temperature, researchers gave several different amounts of alcohol to mice. Then they measured the change in each mouse's body temperature 15 minutes later. In this case, amount of alcohol is the explanatory variable, and change in body temperature is the response variable. When we don't specify the values of either variable but just observe both variables, there may or may not be explanatory and response variables. Whether there are depends on how you plan to use the data.

EXAMPLE Linking SAT Math and Critical Reading Scores

Explanatory or response?

Julie asks, "Can I predict a state's mean SAT Math score if I know its mean SAT Critical Reading score?" Jim wants to know how the mean SAT Math and Critical Reading scores this year in the 50 states are related to each other.

PROBLEM: For each student, identify the explanatory variable and the response variable if possible.

SOLUTION: Julie is treating the mean SAT Critical Reading score as the explanatory variable and the mean SAT Math score as the response variable. Jim is simply interested in exploring the relationship between the two variables. For him, there is no clear explanatory or response variable.

For Practice *Try Exercise* 1

In many studies, the goal is to show that changes in one or more explanatory variables actually *cause* changes in a response variable. However, other explanatory-response relationships don't involve direct causation. In the alcohol and mice study, alcohol actually *causes* a change in body temperature. But there is no cause-and-effect relationship between SAT Math and Critical Reading scores. Because the scores are closely related, we can still use a state's mean SAT Critical Reading score to predict its mean Math score. We will learn how to make such predictions in Section 3.2.

CHECK YOUR UNDERSTANDING

Identify the explanatory and response variables in each setting.

1. How does drinking beer affect the level of alcohol in people's blood? The legal limit for driving in all states is 0.08%. In a study, adult volunteers drank different numbers of cans of beer. Thirty minutes later, a police officer measured their blood alcohol levels.

2. The National Student Loan Survey provides data on the amount of debt for recent college graduates, their current income, and how stressed they feel about college debt. A sociologist looks at the data with the goal of using amount of debt and income to explain the stress caused by college debt.

Displaying Relationships: Scatterplots

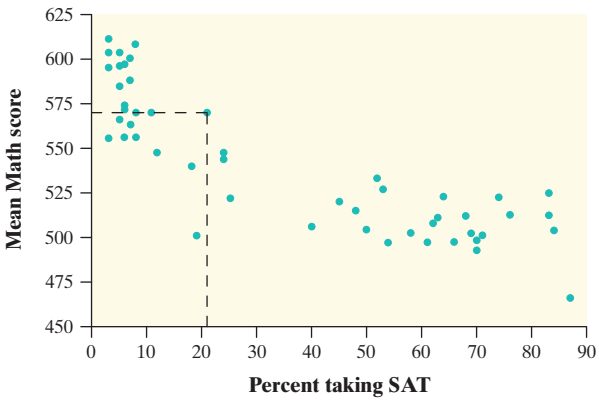


FIGURE 3.2 Scatterplot of the mean SAT Math score in each state against the percent of that state's high school graduates who took the SAT. The dotted lines intersect at the point (21, 570), the values for Colorado.

The most useful graph for displaying the relationship between two quantitative variables is a **scatterplot.** Figure 3.2 shows a scatterplot of the percent of high school graduates in each state who took the SAT and the state's mean SAT Math score in a recent year. We think that "percent taking" will help explain "mean score." So "percent taking" is the explanatory variable and "mean score" is the response variable. We want to see how mean score changes when percent taking changes, so we put percent taking (the explanatory variable) on the horizontal axis. Each point represents a single state. In Colorado, for example, 21% took the SAT, and their mean SAT Math score was 570. Find 21 on the x (horizontal) axis and 570 on the y (vertical) axis. Colorado appears as the point (21, 570).

DEFINITION: Scatterplot

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point in the graph.

Here's a helpful way to remember: the e**X**planatory variable goes on the x axis.

Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y. If there is no explanatory-response distinction, either variable can go on the horizontal axis.

We used computer software to produce Figure 3.2. For some problems, you'll be expected to make scatterplots by hand. Here's how to do it.

HOW TO MAKE A SCATTERPLOT

1. Decide which variable should go on each axis.
2. Label and scale your axes.
3. Plot individual data values.

The following example illustrates the process of constructing a scatterplot.

EXAMPLE SEC Football

Making a scatterplot

At the end of the 2011 college football season, the University of Alabama defeated Louisiana State University for the national championship. Interestingly, both of these teams were from the Southeastern Conference (SEC). Here are the average number of points scored per game and number of wins for each of the twelve teams in the SEC that season.[1]

Team	Alabama	Arkansas	Auburn	Florida	Georgia	Kentucky
Points per game	34.8	36.8	25.7	25.5	32.0	15.8
Wins	12	11	8	7	10	5
Team	Louisiana State	Mississippi	Mississippi State	South Carolina	Tennessee	Vanderbilt
Points per game	35.7	16.1	25.3	30.1	20.3	26.7
Wins	13	2	7	11	5	6

PROBLEM: Make a scatterplot of the relationship between points per game and wins.

SOLUTION: We follow the steps described earlier to make the scatterplot.

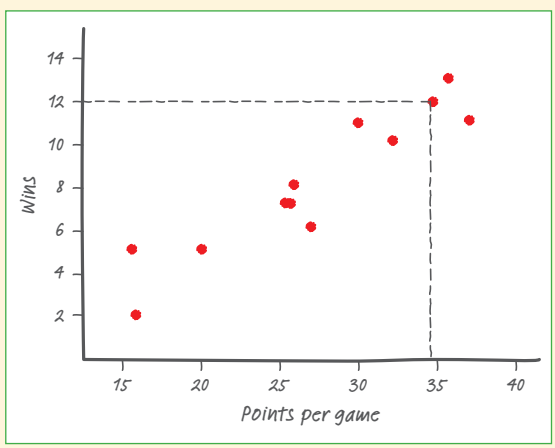


FIGURE 3.3 Completed scatterplot of points per game and wins for the teams in the SEC. The dotted lines intersect at the point (34.8, 12), the values for Alabama.

1. *Decide which variable should go on which axis.* The number of wins a football team has depends on the number of points they score. So we'll use points per game as the explanatory variable (*x* axis) and wins as the response variable (*y* axis).

2. *Label and scale your axes.* We labeled the *x* axis "Points per game" and the *y* axis "Wins." Because the teams' points per game vary from 15.8 to 36.8, we chose a horizontal scale starting at 15 points, with tick marks every 5 points. The teams' wins vary from 2 to 13, so we chose a vertical scale starting at 0 with tick marks every 2 wins.

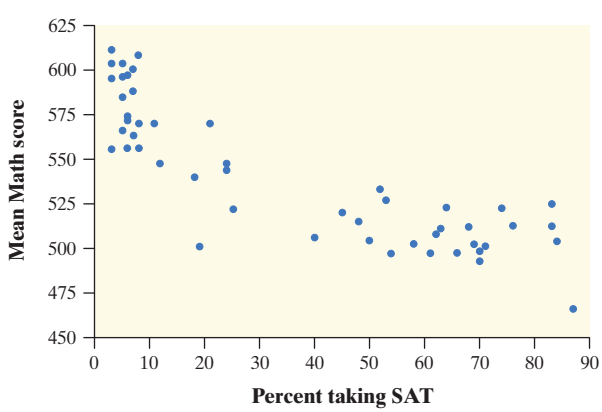
3. *Plot individual data values.* The first team in the table, Alabama, scored 34.8 points per game and had 12 wins. We plot this point directly above 34.8 on the horizontal axis and to the right of 12 on the vertical axis, as shown in Figure 3.3. For the second team in the list, Arkansas, we add the point (36.8, 11) to the graph. By adding the points for the remaining ten teams, we get the completed scatterplot in Figure 3.3.

For Practice *Try Exercise* 5

Describing Scatterplots

To describe a scatterplot, follow the basic strategy of data analysis from Chapters 1 and 2: look for patterns and important departures from those patterns. Let's take a closer look at the scatterplot from Figure 3.2. What do we see?

- The graph shows a clear **direction:** the overall pattern moves from upper left to lower right. That is, states in which higher percents of high school graduates take the SAT tend to have lower mean SAT Math scores. We call this a *negative association* between the two variables.



- The **form** of the relationship is slightly curved. More important, most states fall into one of two distinct *clusters*. In about half of the states, 25% or fewer graduates took the SAT. In the other half, more than 40% took the SAT.
- The **strength** of a relationship in a scatterplot is determined by how closely the points follow a clear form. The overall relationship in Figure 3.2 is moderately strong: states with similar percents taking the SAT tend to have roughly similar mean SAT Math scores.

- Two states stand out in the scatterplot: West Virginia at (19, 501) and Maine at (87, 466). These points can be described as **outliers** because they fall outside the overall pattern.

THINK ABOUT IT

What explains the clusters? There are two widely used college entrance exams, the SAT and the American College Testing (ACT) exam. Each state usually favors one or the other. The ACT states cluster at the left of Figure 3.2 and the SAT states at the right. In ACT states, most students who take the SAT are applying to a selective college that prefers SAT scores. This select group of students has a higher mean score than the much larger group of students who take the SAT in SAT states.

HOW TO EXAMINE A SCATTERPLOT

As in any graph of data, look for the *overall pattern* and for striking *departures* from that pattern.

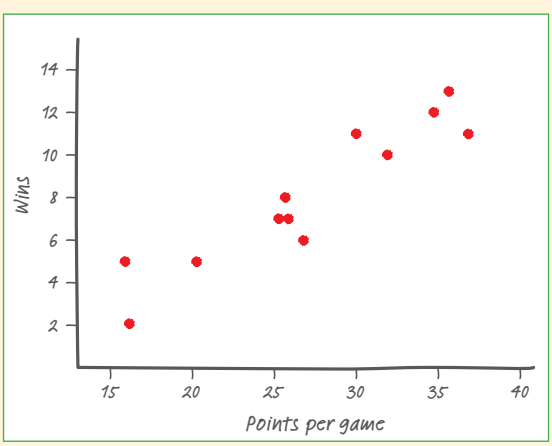
- You can describe the overall pattern of a scatterplot by the **direction, form,** and **strength** of the relationship.
- An important kind of departure is an **outlier,** an individual value that falls outside the overall pattern of the relationship.

Let's practice examining scatterplots using the SEC football data from the previous example.

EXAMPLE SEC Football

Describing a scatterplot

In the last example, we constructed the scatterplot shown below that displays the average number of points scored per game and the number of wins for college football teams in the Southeastern Conference.



PROBLEM: Describe what the scatterplot reveals about the relationship between points per game and wins.

SOLUTION: *Direction:* In general, it appears that teams that score more points per game have more wins and teams that score fewer points per game have fewer wins. We say that there is a *positive association* between points per game and wins.

Form: There seems to be a linear pattern in the graph (that is, the overall pattern follows a straight line).

Strength: Because the points do not vary much from the linear pattern, the relationship is fairly strong. There do not appear to be any values that depart from the linear pattern, so there are no *outliers*.

For Practice *Try Exercise* 7

Even when there is a clear association between two variables in a scatterplot, the direction of the relationship only describes the overall trend—not the relationship for each pair of points. For example, even though teams that score more points per game generally have more wins, Georgia and South Carolina are exceptions to the overall pattern. Georgia scored *more* points per game than South Carolina (32 versus 30.1) but had *fewer* wins (10 versus 11).

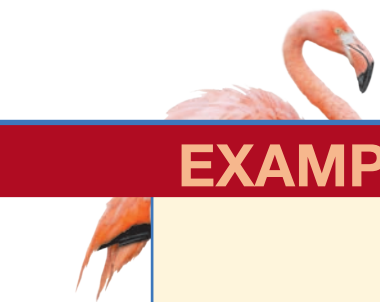
So far, we've seen relationships with two different directions. The number of wins generally increases as the points scored per game increases (**positive association**). The mean SAT score generally goes down as the percent of graduates taking the test increases (**negative association**). Let's give a careful definition for these terms.

DEFINITION: Positive association, negative association

Two variables have a **positive association** when above-average values of one tend to accompany above-average values of the other and when below-average values also tend to occur together.

Two variables have a **negative association** when above-average values of one tend to accompany below-average values of the other.

Of course, not all relationships have a clear direction that we can describe as a positive association or a negative association. Exercise 9 involves a relationship that doesn't have a single direction. This next example, however, illustrates a strong positive association with a simple and important form.



EXAMPLE The Endangered Manatee

Pulling it all together

Manatees are large, gentle, slow-moving creatures found along the coast of Florida. Many manatees are injured or killed by boats. The table below contains data on the number of boats registered in Florida (in thousands) and the number of manatees killed by boats for the years 1977 to 2010.[2]

Florida boat registrations (thousands) and manatees killed by boats

YEAR	BOATS	MANATEES	YEAR	BOATS	MANATEES	YEAR	BOATS	MANATEES
1977	447	13	1989	711	50	2001	944	81
1978	460	21	1990	719	47	2002	962	95
1979	481	24	1991	681	53	2003	978	73
1980	498	16	1992	679	38	2004	983	69
1981	513	24	1993	678	35	2005	1010	79
1982	512	20	1994	696	49	2006	1024	92
1983	526	15	1995	713	42	2007	1027	73
1984	559	34	1996	732	60	2008	1010	90
1985	585	33	1997	755	54	2009	982	97
1986	614	33	1998	809	66	2010	942	83
1987	645	39	1999	830	82			
1988	675	43	2000	880	78			

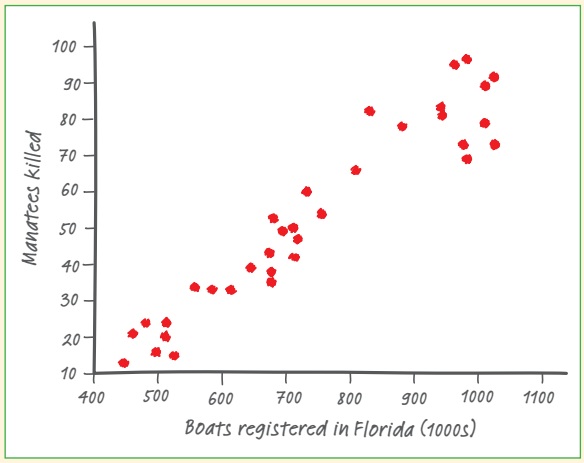
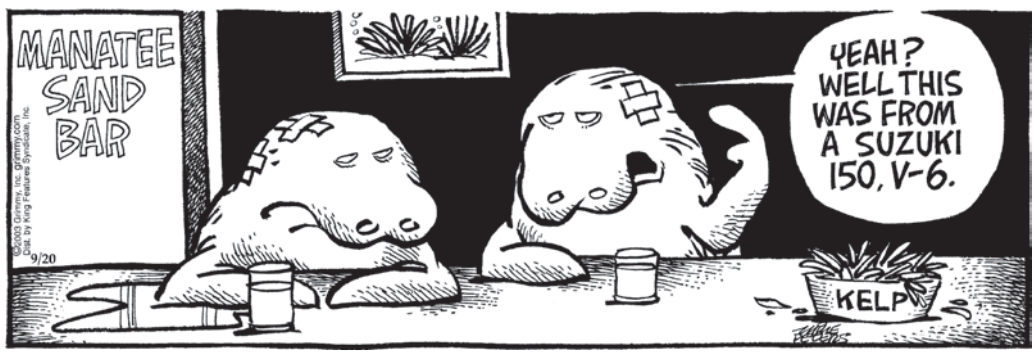


FIGURE 3.4 Scatterplot of the number of Florida manatees killed by boats from 1977 to 2010 against the number of boats registered in Florida that year.

PROBLEM: Make a scatterplot to show the relationship between the number of manatees killed and the number of registered boats. Describe what you see.

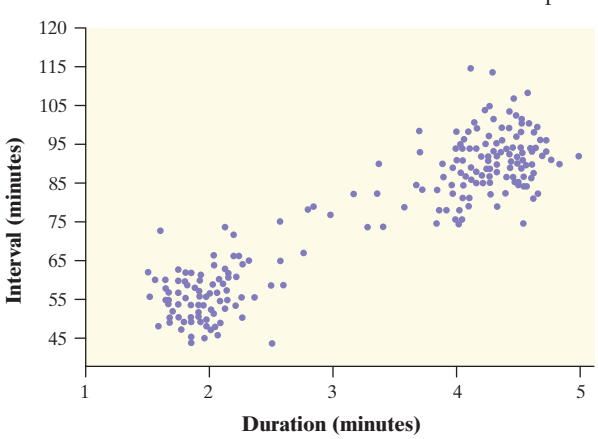
SOLUTION: For the scatterplot, we'll use "boats registered" as the explanatory variable and "manatees killed" as the response variable. Figure 3.4 is our completed scatterplot. There is a positive association—more boats registered goes with more manatees killed. The form of the relationship is linear. That is, the overall pattern follows a straight line from lower left to upper right. The relationship is strong because the points don't deviate greatly from a line, except for the 4 years that have a high number of boats registered, but fewer deaths than expected based on the linear pattern.

For Practice *Try Exercise* 13



CHECK YOUR UNDERSTANDING

In the chapter-opening Case Study (page 141), the Starnes family arrived at Old Faithful after it had erupted. They wondered how long it would be until the next eruption. Here is a scatterplot that plots the interval between consecutive eruptions of Old Faithful against the duration of the previous eruption, for the month prior to their visit.



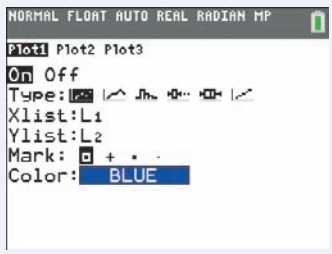
1. Describe the direction of the relationship. Explain why this makes sense.
2. What form does the relationship take? Why are there two clusters of points?
3. How strong is the relationship? Justify your answer.
4. Are there any outliers?
5. What information does the Starnes family need to predict when the next eruption will occur?

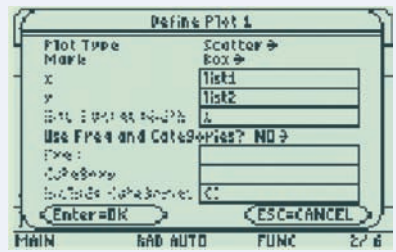
7. TECHNOLOGY CORNER SCATTERPLOTS ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

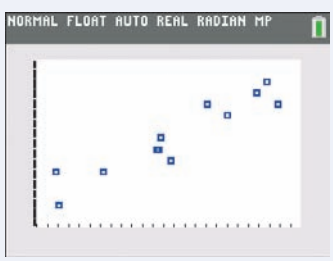
Making scatterplots with technology is much easier than constructing them by hand. We'll use the SEC football data from page 146 to show how to construct a scatterplot on a TI-83/84 or TI-89.

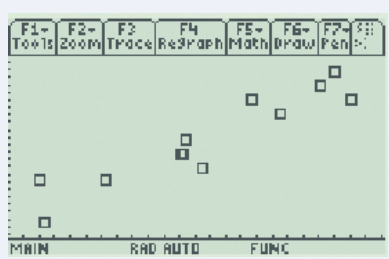
- Enter the data values into your lists. Put the points per game in L1/list1 and the number of wins in L2/list2.
- Define a scatterplot in the statistics plot menu (press F2 on the TI-89). Specify the settings shown below.





- Use ZoomStat (ZoomData on the TI-89) to obtain a graph. The calculator will set the window dimensions automatically by looking at the values in L1/list1 and L2/list2.





Notice that there are no scales on the axes and that the axes are not labeled. If you copy a scatterplot from your calculator onto your paper, make sure that you scale and label the axes.

AP® EXAM TIP If you are asked to make a scatterplot on a free-response question, be sure to label and scale both axes. *Don't* just copy an unlabeled calculator graph directly onto your paper.

Measuring Linear Association: Correlation

A scatterplot displays the direction, form, and strength of the relationship between two quantitative variables. Linear relationships are particularly important because a straight line is a simple pattern that is quite common. A linear relationship is strong if the points lie close to a straight line and weak if they are widely scattered about a line. Unfortunately, *our eyes are not good judges of how strong a linear relationship is.* The two scatterplots in Figure 3.5 (on the facing page) show the same data, but the graph on the right is drawn smaller in a large field. The right-hand graph seems to show a stronger linear relationship.

Because it's easy to be fooled by different scales or by the amount of space around the cloud of points in a scatterplot, we need to use a numerical measure to supplement the graph. **Correlation** is the measure we use.

Some people refer to *r* as the "correlation coefficient."

DEFINITION: Correlation *r*

The **correlation *r*** measures the direction and strength of the linear relationship between two quantitative variables.

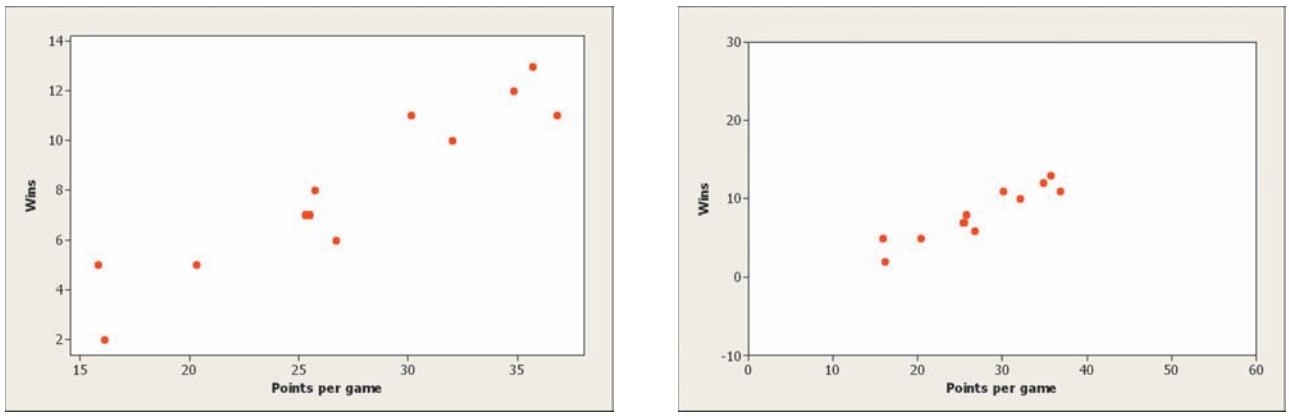


FIGURE 3.5 Two Minitab scatterplots of the same data. The straight-line pattern in the graph on the right appears stronger because of the surrounding space.

How good are you at estimating the correlation by eye from a scatterplot? To find out, try an online applet. Just search for "guess the correlation applets."

The correlation r is always a number between −1 and 1. Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association. Values of r near 0 indicate a very weak linear relationship. The strength of the linear relationship increases as r moves away from 0 toward either -1 or 1. The extreme values $r = -1$ and $r = 1$ occur *only* in the case of a perfect linear relationship, when the points lie exactly along a straight line.

Figure 3.6 shows scatterplots that correspond to various values of r. To make the meaning of r clearer, the standard deviations of both variables in these plots are equal, and the horizontal and vertical scales are the same. The correlation describes the direction and strength of the linear relationship in each graph.

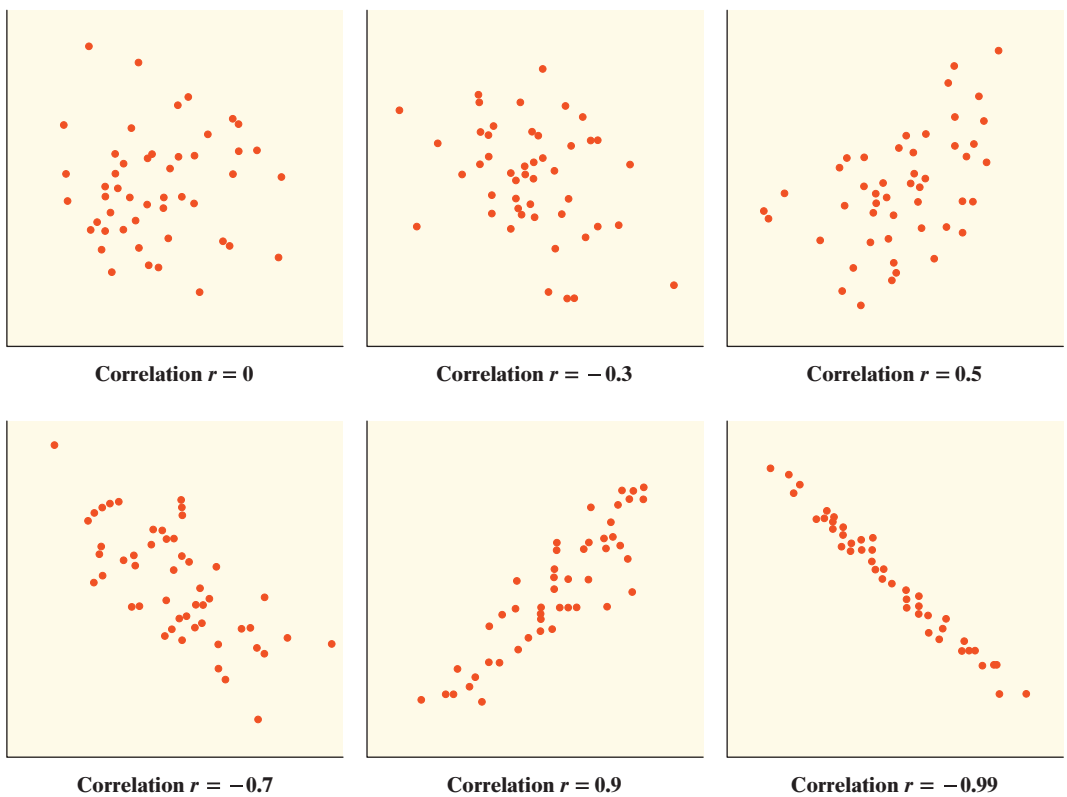


FIGURE 3.6 How correlation measures the strength of a linear relationship. Patterns closer to a straight line have correlations closer to 1 or −1.

The following Activity lets you explore some important properties of the correlation.

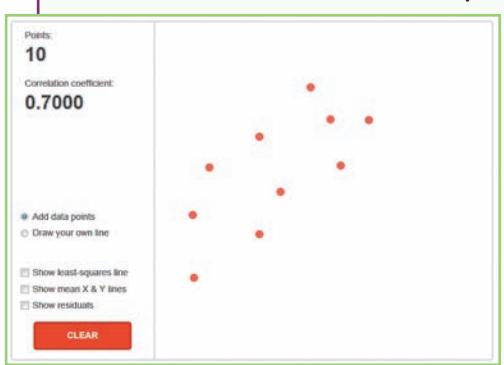
ACTIVITY | *Correlation and Regression* applet

MATERIALS:
Computer with Internet connection

Go to the book's Web site, www.whfreeman.com/tps5e, and launch the *Correlation and Regression* applet.

1. You are going to use the *Correlation and Regression* applet to make several scatterplots with 10 points that have correlation close to 0.7.

(a) Start by putting two points on the graph. What's the value of the correlation? Why does this make sense?



(b) Make a lower-left to upper-right pattern of 10 points with correlation about $r = 0.7$. (You can drag points up or down to adjust r after you have 10 points.)

(c) Make another scatterplot: this one should have 9 points in a vertical stack at the left of the plot. Add 1 point far to the right and move it until the correlation is close to 0.7.

(d) Make a third scatterplot: make this one with 10 points in a curved pattern that starts at the lower left, rises to the right, then falls again at the far right. Adjust the points up or down until you have a very smooth curve with correlation close to 0.7.

Summarize: If you know that the correlation between two variables is $r = 0.7$, what can you say about the form of the relationship?

2. Click on the scatterplot to create a group of 10 points in the lower-left corner of the scatterplot with a strong straight-line pattern (correlation about 0.9).

(a) Add 1 point at the upper right that is in line with the first 10. How does the correlation change?

(b) Drag this last point straight down. How small can you make the correlation? Can you make the correlation negative?

Summarize: What did you learn from Step 2 about the effect of a single point on the correlation?

Now that you know what information the correlation provides—and doesn't provide—let's look at an example that shows how to interpret it.

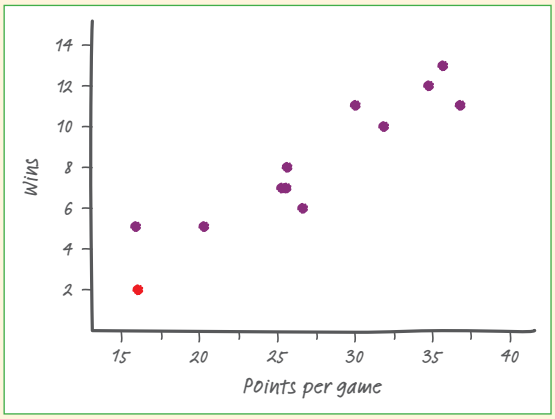
EXAMPLE SEC Football

Interpreting correlation

PROBLEM: Our earlier scatterplot of the average points per game and number of wins for college football teams in the SEC is repeated at top right. For these data, $r = 0.936$.

(a) Interpret the value of r in context.

(b) The point highlighted in red on the scatterplot is Mississippi. What effect does Mississippi have on the correlation? Justify your answer.



SOLUTION:

(a) The correlation of 0.936 confirms what we see in the scatterplot: there is a strong, positive linear relationship between points per game and wins in the SEC.

(b) Mississippi makes the correlation closer to 1 (stronger). If Mississippi were not included, the remaining points wouldn't be as tightly clustered in a linear pattern.

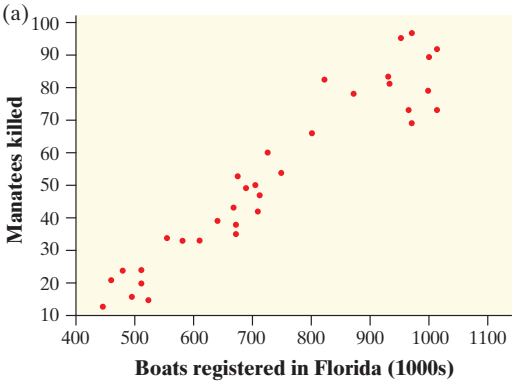
For Practice *Try Exercise* 21

AP® EXAM TIP If you're asked to interpret a correlation, start by looking at a scatterplot of the data. Then be sure to address direction, form, strength, and outliers (sound familiar?) and put your answer in context.

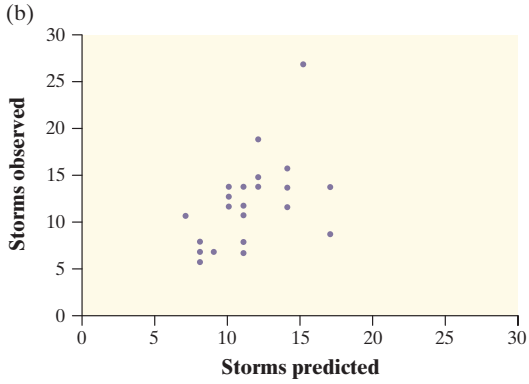
CHECK YOUR UNDERSTANDING

The scatterplots below show four sets of real data: (a) repeats the manatee plot in Figure 3.4 (page 149); (b) shows the number of named tropical storms and the number predicted before the start of hurricane season each year between 1984 and 2007 by William Gray of Colorado State University; (c) plots the healing rate in micrometers (millionths of a meter) per hour for the two front limbs of several newts in an experiment; and (d) shows stock market performance in consecutive years over a 56-year period. For each graph, estimate the correlation r. Then interpret the value of r in context.

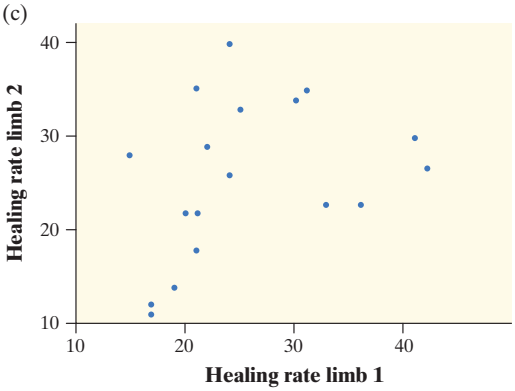
(a)



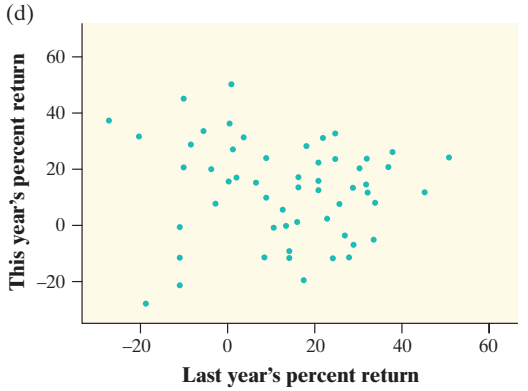
(b)



(c)



(d)



Calculating Correlation Now that you have some idea of how to interpret the correlation, let's look at how it's calculated.

HOW TO CALCULATE THE CORRELATION *r*

Suppose that we have data on variables x and y for n individuals. The values for the first individual are x_1 and y_1, the values for the second individual are x_2 and y_2, and so on. The means and standard deviations of the two variables are $\bar{x}$ and s_x for the x-values, and $\bar{y}$ and s_y for the y-values. The correlation r between x and y is

$$r = \frac{1}{n-1}\left[\left(\frac{x_1-\bar{x}}{s_x}\right)\left(\frac{y_1-\bar{y}}{s_y}\right) + \left(\frac{x_2-\bar{x}}{s_x}\right)\left(\frac{y_2-\bar{y}}{s_y}\right) + \cdots + \left(\frac{x_n-\bar{x}}{s_x}\right)\left(\frac{y_n-\bar{y}}{s_y}\right)\right]$$

or, more compactly,

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

The formula for the correlation r is a bit complex. It helps us see what correlation is, but in practice, you should use your calculator or software to find r. Exercises 19 and 20 ask you to calculate a correlation step-by-step from the definition to solidify its meaning.

The formula for r begins by standardizing the observations. Let's use the familiar SEC football data to perform the required calculations. The table below shows the values of points per game x and number of wins y for the SEC college football teams. For these data, $\bar{x} = 27.07$ and $s_x = 7.16$.

Team	Alabama	Arkansas	Auburn	Florida	Georgia	Kentucky
Points per game	34.8	36.8	25.7	25.5	32.0	15.8
Wins	12	11	8	7	10	5
Team	Louisiana State	Mississippi	Mississippi State	South Carolina	Tennessee	Vanderbilt
Points per game	35.7	16.1	25.3	30.1	20.3	26.7
Wins	13	2	7	11	5	6

The value

$$\frac{x_i-\bar{x}}{s_x}$$

in the correlation formula is the standardized points per game (z-score) of the ith team. For the first team in the table (Alabama), the corresponding z-score is

$$z_x = \frac{34.8 - 27.07}{7.16} = 1.08$$

That is, Alabama's points per game total (34.8) is a little more than 1 standard deviation above the mean points per game for the SEC teams.

Standardized values have no units—in this example, they are no longer measured in points.

To standardize the number of wins, we use $\bar{y} = 8.08$ and $s_y = 3.34$. For Alabama, $z_y = \dfrac{12 - 8.08}{3.34} = 1.17$. Alabama's number of wins (12) is 1.17 standard deviations above the mean number of wins for SEC teams. When we multiply this team's two z-scores, we get a product of 1.2636. The correlation r is an "average" of the products of the standardized scores for all the teams. Just as in the case of the standard deviation s_x, the average here divides by one fewer than the number of individuals. Finishing the calculation reveals that $r = 0.936$ for the SEC teams.

Some people like to write the correlation formula as

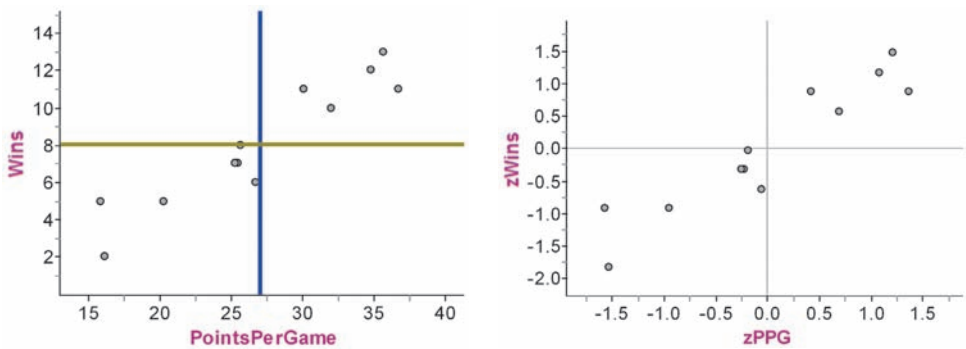
$$r = \frac{1}{n-1}\sum z_x z_y$$

to emphasize the product of standardized scores in the calculation.

THINK ABOUT IT

What does correlation measure? The Fathom screen shots below provide more detail. At the left is a scatterplot of the SEC football data with two lines added—a vertical line at the group's mean points per game and a horizontal line at the mean number of wins of the group. Most of the points fall in the upper-right or lower-left "quadrants" of the graph. That is, teams with above-average points per game tend to have above-average numbers of wins, and teams with below-average points per game tend to have numbers of wins that are below average. This confirms the positive association between the variables.

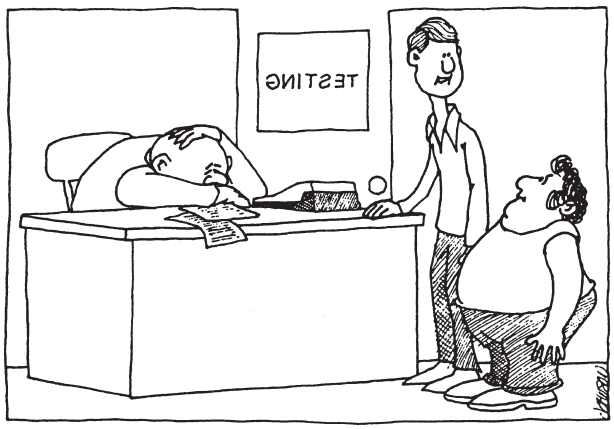
Below on the right is a scatterplot of the standardized scores. To get this graph, we transformed both the x- and the y-values by subtracting their mean and dividing by their standard deviation. As we saw in Chapter 2, standardizing a data set converts the mean to 0 and the standard deviation to 1. That's why the vertical and horizontal lines in the right-hand graph are both at 0.



Notice that all the products of the standardized values will be positive—not surprising, considering the strong positive association between the variables. What if there was a negative association between two variables? Most of the points would be in the upper-left and lower-right "quadrants" and their z-score products would be negative, resulting in a negative correlation.

Facts about Correlation

How correlation behaves is more important than the details of the formula. Here's what you need to know in order to interpret correlation correctly.



"He says we've ruined his positive correlation between height and weight."

1. *Correlation makes no distinction between explanatory and response variables.* It makes no difference which variable you call x and which you call y in calculating the correlation. Can you see why from the formula?

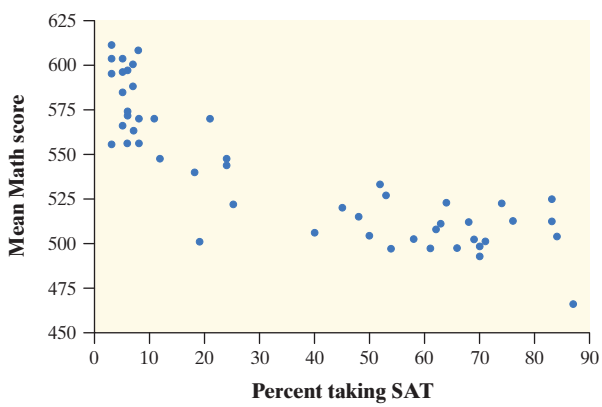
$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

2. Because r uses the standardized values of the observations, *r does not change when we change the units of measurement of x, y, or both.* Measuring height in centimeters rather than inches and weight in kilograms rather than pounds does not change the correlation between height and weight.

3. *The correlation r itself has no unit of measurement.* It is just a number.

Describing the relationship between two variables is more complex than describing the distribution of one variable. Here are some cautions to keep in mind when you use correlation.

- *Correlation does not imply causation.* Even when a scatterplot shows a strong linear relationship between two variables, we can't conclude that changes in one variable cause changes in the other. For example, looking at data from the last 10 years, there is a strong positive relationship between the number of high school students who own a cell phone and the number of students who pass the AP® Statistics exam. Does this mean that buying a cell phone will help you pass the AP® exam? Not likely. Instead, the correlation is positive because both of these variables are increasing over time.
- *Correlation requires that both variables be quantitative,* so that it makes sense to do the arithmetic indicated by the formula for r. We cannot calculate a correlation between the incomes of a group of people and what city they live in because city is a categorical variable.
- Correlation only measures the strength of a linear relationship between two variables. *Correlation does not describe curved relationships between variables, no matter how strong the relationship is.* A correlation of 0 doesn't guarantee that there's *no* relationship between two variables, just that there's no *linear* relationship.
- *A value of r close to 1 or -1 does not guarantee a linear relationship between two variables.* A scatterplot with a clear curved form can have a correlation that is close to 1 or -1. For example, the correlation between percent taking the SAT and mean Math score is close to -1, but the association is clearly curved. Always plot your data!



- *Like the mean and standard deviation, the correlation is not resistant: r is strongly affected by a few outlying observations.* Use r with caution when outliers appear in the scatterplot.
- *Correlation is not a complete summary of two-variable data,* even when the relationship between the variables is linear. You should give the means and standard deviations of both x and y along with the correlation.

Of course, even giving means, standard deviations, and the correlation for "state SAT Math scores" and "percent taking" will not point out the clusters in Figure 3.2. Numerical summaries complement plots of data, but they do not replace them.

EXAMPLE Scoring Figure Skaters

Why correlation doesn't tell the whole story

Until a scandal at the 2002 Olympics brought change, figure skating was scored by judges on a scale from 0.0 to 6.0. The scores were often controversial. We have the scores awarded by two judges, Pierre and Elena, for many skaters. How well do they agree? We calculate that the correlation between their scores is $r = 0.9$. But the mean of Pierre's scores is 0.8 point lower than Elena's mean.

These facts don't contradict each other. They simply give different kinds of information. The mean scores show that Pierre awards lower scores than Elena. But because Pierre gives *every* skater a score about 0.8 point lower than Elena does, the correlation remains high. Adding the same number to all values of either x or y does not change the correlation. If both judges score the same skaters, the competition is scored consistently because Pierre and Elena agree on which performances are better than others. The high r shows their agreement. But if Pierre scores some skaters and Elena others, we should add 0.8 point to Pierre's scores to arrive at a fair comparison.

DATA EXPLORATION The SAT essay: Is longer better?

Following the debut of the new SAT Writing test in March 2005, Dr. Les Perelman from the Massachusetts Institute of Technology stirred controversy by reporting, "It appeared to me that regardless of what a student wrote, the longer the essay, the higher the score." He went on to say, "I have never found a quantifiable predictor in 25 years of grading that was anywhere as strong as this one. If you just graded them based on length without ever reading them, you'd be right over 90 percent of the time."[3] The table below shows the data that Dr. Perelman used to draw his conclusions.[4]

Length of essay and score for a sample of SAT essays

Words:	460	422	402	365	357	278	236	201	168	156	133
Score:	6	6	5	5	6	5	4	4	4	3	2
Words:	114	108	100	403	401	388	320	258	236	189	128
Score:	2	1	1	5	6	6	5	4	4	3	2
Words:	67	697	387	355	337	325	272	150	135		
Score:	1	6	6	5	5	4	4	2	3		

Does this mean that if students write a lot, they are guaranteed high scores? Carry out your own analysis of the data. How would you respond to each of Dr. Perelman's claims?

Section 3.1 Summary

- A **scatterplot** displays the relationship between two quantitative variables measured on the same individuals. Mark values of one variable on the horizontal axis (x axis) and values of the other variable on the vertical axis (y axis). Plot each individual's data as a point on the graph.
- If we think that a variable x may help explain, predict, or even cause changes in another variable y, we call x an **explanatory variable** and y a **response variable.** Always plot the explanatory variable, if there is one, on the x axis of a scatterplot. Plot the response variable on the y axis.
- In examining a scatterplot, look for an overall pattern showing the **direction, form,** and **strength** of the relationship and then look for **outliers** or other departures from this pattern.
- **Direction:** If the relationship has a clear direction, we speak of either **positive association** (above-average values of the two variables tend to occur together) or **negative association** (above-average values of one variable tend to occur with below-average values of the other variable).
- **Form:** Linear relationships, where the points show a straight-line pattern, are an important form of relationship between two variables. Curved relationships and clusters are other forms to watch for.
- **Strength:** The strength of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.
- The **correlation** r measures the strength and direction of the linear association between two quantitative variables x and y. Although you can calculate a correlation for any scatterplot, r measures strength for only straight-line relationships.
- Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association. Correlation always satisfies $-1 \le r \le 1$ and indicates the strength of a linear relationship by how close it is to -1 or 1. Perfect correlation, $r = \pm 1$, occurs only when the points on a scatterplot lie exactly on a straight line.
- Remember these important facts about r: Correlation does not imply causation. Correlation ignores the distinction between explanatory and response variables. The value of r is not affected by changes in the unit of measurement of either variable. Correlation is not resistant, so outliers can greatly change the value of r.

3.1 TECHNOLOGY CORNER

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

7. Scatterplots on the calculator page 150

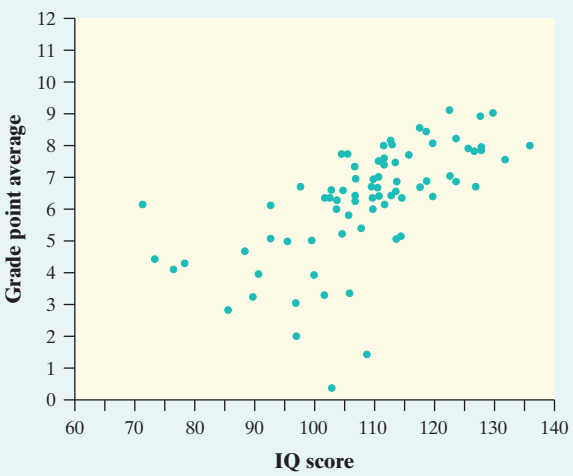
Section 3.1 Exercises

1. **Coral reefs** How sensitive to changes in water temperature are coral reefs? To find out, measure the growth of corals in aquariums where the water temperature is controlled at different levels. Growth is measured by weighing the coral before and after the experiment. What are the explanatory and response variables? Are they categorical or quantitative?

pg 144

2. **Treating breast cancer** Early on, the most common treatment for breast cancer was removal of the breast. It is now usual to remove only the tumor and nearby lymph nodes, followed by radiation. The change in policy was due to a large medical experiment that compared the two treatments. Some breast cancer patients, chosen at random, were given one or the other treatment. The patients were closely followed to see how long they lived following surgery. What are the explanatory and response variables? Are they categorical or quantitative?

3. **IQ and grades** Do students with higher IQ test scores tend to do better in school? The figure below shows a scatterplot of IQ and school grade point average (GPA) for all 78 seventh-grade students in a rural midwestern school. (GPA was recorded on a 12-point scale with A+ = 12, A = 11, A− = 10, B+ = 9, . . . , D− = 1, and F = 0.)[5]

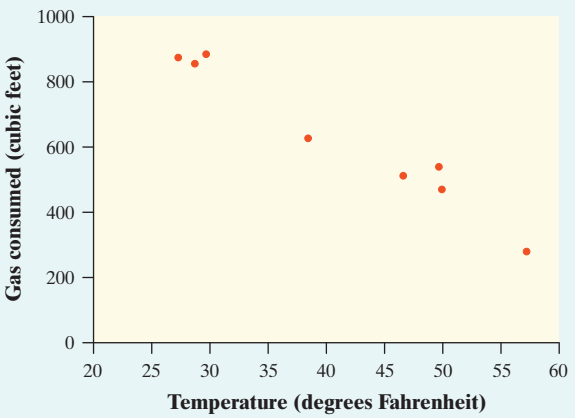


(a) Does the plot show a positive or negative association between the variables? Why does this make sense?

(b) What is the form of the relationship? Is it very strong? Explain your answers.

(c) At the bottom of the plot are several points that we might call outliers. One student in particular has a very low GPA despite an average IQ score. What are the approximate IQ and GPA for this student?

4. **How much gas?** Joan is concerned about the amount of energy she uses to heat her home. The graph below plots the mean number of cubic feet of gas per day that Joan used each month against the average temperature that month (in degrees Fahrenheit) for one heating season.



(a) Does the plot show a positive or negative association between the variables? Why does this make sense?

(b) What is the form of the relationship? Is it very strong? Explain your answers.

(c) Explain what the point at the bottom right of the plot represents.

5. **Heavy backpacks** Ninth-grade students at the Webb Schools go on a backpacking trip each fall. Students are divided into hiking groups of size 8 by selecting names from a hat. Before leaving, students and their backpacks are weighed. The data here are from one hiking group in a recent year. Make a scatterplot by hand that shows how backpack weight relates to body weight.

pg 145

Body weight (lb):	120	187	109	103	131	165	158	116
Backpack weight (lb):	26	30	26	24	29	35	31	28

6. **Bird colonies** One of nature's patterns connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 13 colonies of sparrowhawks:[6]

Percent return:	74	66	81	52	73	62	52	45	62	46	60	46	38
New adults:	5	6	8	11	12	15	16	17	18	18	19	20	20

Make a scatterplot by hand that shows how the number of new adults relates to the percent of returning birds.

pg 147

7. **Heavy backpacks** Refer to your graph from Exercise 5.

(a) Describe the relationship between body weight and backpack weight for this group of hikers.

(b) One of the hikers is a possible outlier. Identify the body weight and backpack weight for this hiker. How does this hiker affect the form of the association?

8. **Bird colonies** Refer to your graph from Exercise 6.

(a) Describe the relationship between number of new sparrowhawks in a colony and percent of returning adults.

(b) For short-lived birds, the association between these variables is positive: changes in weather and food supply drive the populations of new and returning birds up or down together. For long-lived territorial birds, on the other hand, the association is negative because returning birds claim their territories in the colony and don't leave room for new recruits. Which type of species is the sparrowhawk? Explain.

9. **Does fast driving waste fuel?** How does the fuel consumption of a car change as its speed increases? Here are data for a British Ford Escort. Speed is measured in kilometers per hour, and fuel consumption is measured in liters of gasoline used per 100 kilometers traveled.[7]

Speed (km/h)	Fuel used (liters/100 km)	Speed (km/h)	Fuel used (liters/100 km)
10	21.00	90	7.57
20	13.00	100	8.27
30	10.00	110	9.03
40	8.00	120	9.87
50	7.00	130	10.79
60	5.90	140	11.77
70	6.30	150	12.83
80	6.95		

(a) Use your calculator to help sketch a scatterplot.

(b) Describe the form of the relationship. Why is it not linear? Explain why the form of the relationship makes sense.

(c) It does not make sense to describe the variables as either positively associated or negatively associated. Why?

(d) Is the relationship reasonably strong or quite weak? Explain your answer.

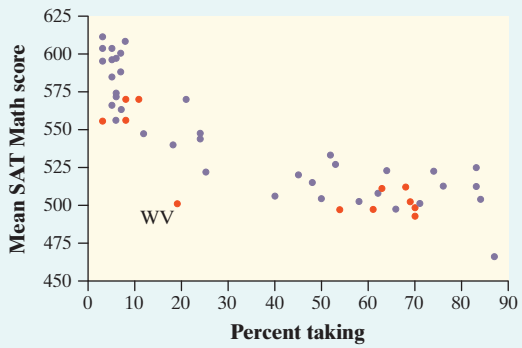
10. **Do heavier people burn more energy?** Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. We have data on the lean body mass and resting metabolic rate for 12 women who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours. The researchers believe that lean body mass is an important influence on metabolic rate.

Mass:	36.1	54.6	48.5	42.0	50.6	42.0	40.3	33.1	42.4	34.5	51.1	41.2
Rate:	995	1425	1396	1418	1502	1256	1189	913	1124	1052	1347	1204

(a) Use your calculator to help sketch a scatterplot to examine the researchers' belief.

(b) Describe the direction, form, and strength of the relationship.

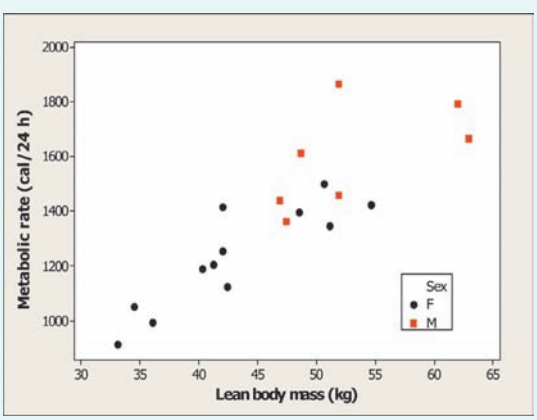
11. **Southern education** For a long time, the South has lagged behind the rest of the United States in the performance of its schools. Efforts to improve education have reduced the gap. We wonder if the South stands out in our study of state average SAT Math scores. The figure below enhances the scatterplot in Figure 3.2 (page 145) by plotting 12 southern states in red.



(a) What does the graph suggest about the southern states?

(b) The point for West Virginia is labeled in the graph. Explain how this state is an outlier.

12. **Do heavier people burn more energy?** The study of dieting described in Exercise 10 collected data on the lean body mass (in kilograms) and metabolic rate (in calories) for 12 female and 7 male subjects. The figure below is a scatterplot of the data for all 19 subjects, with separate symbols for males and females.



Does the same overall pattern hold for both women and men? What difference between the sexes do you see from the graph?

pg 148

13. **Merlins breeding** The percent of an animal species in the wild that survives to breed again is often lower following a successful breeding season. A study of

merlins (small falcons) in northern Sweden observed the number of breeding pairs in an isolated area and the percent of males (banded for identification) that returned the next breeding season. Here are data for seven years:[8]

Breeding pairs:	28	29	29	29	30	32	33
Percent return:	82	83	70	61	69	58	43

Make a scatterplot to display the relationship between breeding pairs and percent return. Describe what you see.

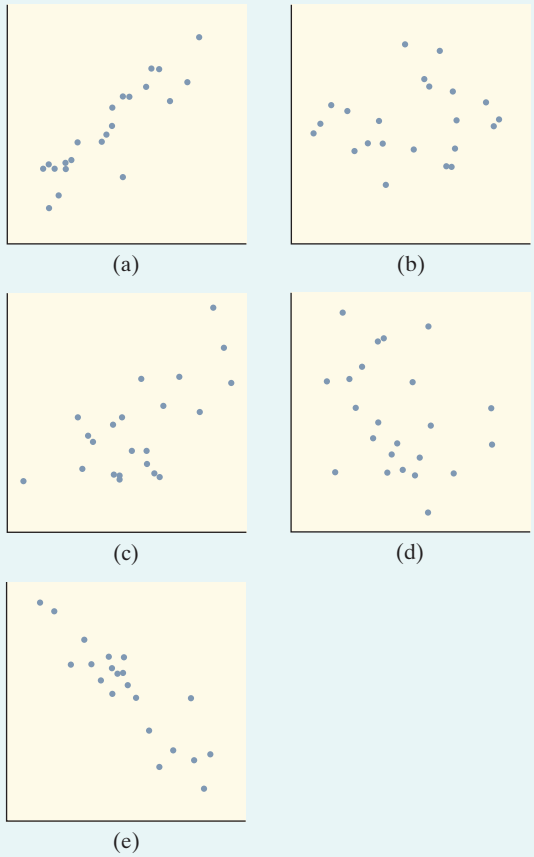
14. **Does social rejection hurt?** We often describe our emotional reaction to social rejection as "pain." Does social rejection cause activity in areas of the brain that are known to be activated by physical pain? If it does, we really do experience social and physical pain in similar ways. Psychologists first included and then deliberately excluded individuals from a social activity while they measured changes in brain activity. After each activity, the subjects filled out questionnaires that assessed how excluded they felt. The table below shows data for 13 subjects.[9] "Social distress" is measured by each subject's questionnaire score after exclusion relative to the score after inclusion. (So values greater than 1 show the degree of distress caused by exclusion.) "Brain activity" is the change in activity in a region of the brain that is activated by physical pain. (So positive values show more pain.)

Subject	Social distress	Brain activity
1	1.26	−0.055
2	1.85	−0.040
3	1.10	−0.026
4	2.50	−0.017
5	2.17	−0.017
6	2.67	0.017
7	2.01	0.021
8	2.18	0.025
9	2.58	0.027
10	2.75	0.033
11	2.75	0.064
12	3.33	0.077
13	3.65	0.124

Make a scatterplot to display the relationship between social distress and brain activity. Describe what you see.

15. **Matching correlations** Match each of the following scatterplots to the r below that best describes it. (Some r's will be left over.)

$$r = -0.9 \quad r = -0.7 \quad r = -0.3 \quad r = 0$$
$$r = 0.3 \quad r = 0.7 \quad r = 0.9$$



16. **Rank the correlations** Consider each of the following relationships: the heights of fathers and the heights of their adult sons, the heights of husbands and the heights of their wives, and the heights of women at age 4 and their heights at age 18. Rank the correlations between these pairs of variables from largest to smallest. Explain your reasoning.

17. **Correlation blunders** Each of the following statements contains an error. Explain what's wrong in each case.

(a) "There is a high correlation between the gender of American workers and their income."

(b) "We found a high correlation ($r = 1.09$) between students' ratings of faculty teaching and ratings made by other faculty members."

(c) "The correlation between planting rate and yield of corn was found to be $r = 0.23$ bushel."

18. **Teaching and research** A college newspaper interviews a psychologist about student ratings of the teaching of faculty members. The psychologist says, "The evidence indicates that the correlation between the research productivity and teaching rating of faculty members is close to zero." The paper reports this as "Professor McDaniel said that good researchers tend to be poor teachers, and vice versa." Explain why the paper's report is wrong. Write a statement in plain language (don't use the word "correlation") to explain the psychologist's meaning.

19. **Dem bones** Archaeopteryx is an extinct beast having feathers like a bird but teeth and a long bony tail like a reptile. Only six fossil specimens are known. Because these specimens differ greatly in size, some scientists think they are different species rather than individuals from the same species. We will examine some data. If the specimens belong to the same species and differ in size because some are younger than others, there should be a positive linear relationship between the lengths of a pair of bones from all individuals. An outlier from this relationship would suggest a different species. Here are data on the lengths in centimeters of the femur (a leg bone) and the humerus (a bone in the upper arm) for the five specimens that preserve both bones:[10]

Femur (x):	38	56	59	64	74
Humerus (y):	41	63	70	72	84

(a) Make a scatterplot. Do you think that all five specimens come from the same species? Explain.

(b) Find the correlation r step by step, using the formula on page 154. Explain how your value for r matches your graph in part (a).

20. **Data on dating** A student wonders if tall women tend to date taller men than do short women. She measures herself, her dormitory roommate, and the women in the adjoining rooms. Then she measures the next man each woman dates. Here are the data (heights in inches):

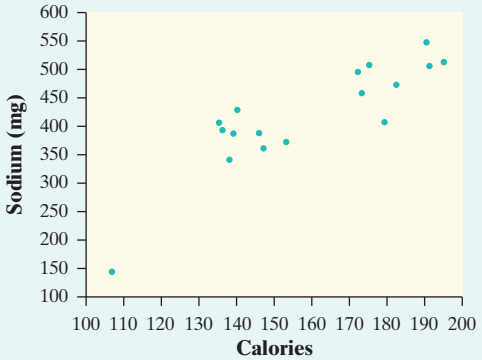
Women (x):	66	64	66	65	70	65
Men (y):	72	68	70	68	71	65

(a) Make a scatterplot of these data. Based on the scatterplot, do you expect the correlation to be positive or negative? Near ± 1 or not?

(b) Find the correlation r step by step, using the formula on page 154. Do the data show that taller women tend to date taller men?

21. **Hot dogs** Are hot dogs that are high in calories also high in salt? The figure below is a scatterplot of the calories and salt content (measured as milligrams of sodium) in 17 brands of meat hot dogs.[11]

pg 152



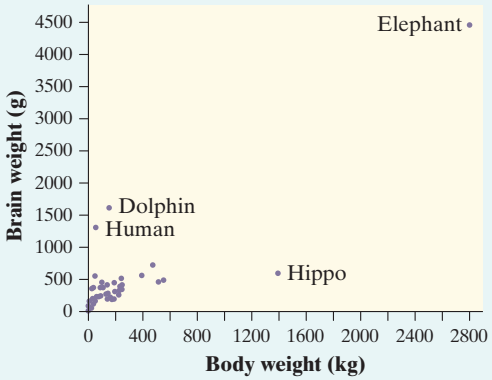
(a) The correlation for these data is $r = 0.87$. Explain what this value means.

(b) What effect does the hot dog brand with the lowest calorie content have on the correlation? Justify your answer.

22. **All brawn?** The figure below plots the average brain weight in grams versus average body weight in kilograms for 96 species of mammals.[12] There are many small mammals whose points overlap at the lower left.

(a) The correlation between body weight and brain weight is $r = 0.86$. Explain what this value means.

(b) What effect does the elephant have on the correlation? Justify your answer.



23. **Dem bones** Refer to Exercise 19.

(a) How would r change if the bones had been measured in millimeters instead of centimeters? (There are 10 millimeters in a centimeter.)

(b) If the x and y variables are reversed, how would the correlation change? Explain.

24. **Data on dating** Refer to Exercise 20.

(a) How would r change if all the men were 6 inches shorter than the heights given in the table? Does the correlation tell us if women tend to date men taller than themselves?

(b) If heights were measured in centimeters rather than inches, how would the correlation change? (There are 2.54 centimeters in an inch.)

25. **Strong association but no correlation** The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose that this relationship is very regular, as shown by the following data on speed (miles per hour) and mileage (miles per gallon).

Speed:	20	30	40	50	60
Mileage:	24	28	30	28	24

(a) Make a scatterplot to show the relationship between speed and mileage.

(b) Calculate the correlation for these data by hand or using technology.

(c) Explain why the correlation has the value found in part (b) even though there is a strong relationship between speed and mileage.

26. What affects correlation? Here are some hypothetical data:

x:	1	2	3	4	10	10
y:	1	3	3	5	1	11

(a) Make a scatterplot to show the relationship between x and y.

(b) Calculate the correlation for these data by hand or using technology.

(c) What is responsible for reducing the correlation to the value in part (b) despite a strong straight-line relationship between x and y in most of the observations?

Multiple choice: Select the best answer for Exercises 27 to 32.

27. You have data for many years on the average price of a barrel of oil and the average retail price of a gallon of unleaded regular gasoline. If you want to see how well the price of oil predicts the price of gas, then you should make a scatterplot with _____ as the explanatory variable.

(a) the price of oil (c) the year (e) time

(b) the price of gas (d) either oil price or gas price

28. In a scatterplot of the average price of a barrel of oil and the average retail price of a gallon of gas, you expect to see

(a) very little association.

(b) a weak negative association.

(c) a strong negative association.

(d) a weak positive association.

(e) a strong positive association.

29. The following graph plots the gas mileage (miles per gallon) of various cars from the same model year versus the weight of these cars in thousands of pounds. The points marked with red dots correspond to cars made in Japan. From this plot, we may conclude that

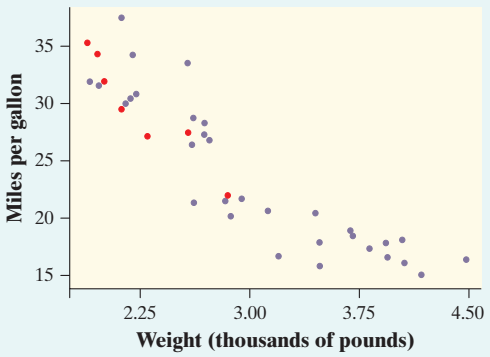
(a) there is a positive association between weight and gas mileage for Japanese cars.

(b) the correlation between weight and gas mileage for all the cars is close to 1.

(c) there is little difference between Japanese cars and cars made in other countries.

(d) Japanese cars tend to be lighter in weight than other cars.

(e) Japanese cars tend to get worse gas mileage than other cars.



30. If women always married men who were 2 years older than themselves, what would the correlation between the ages of husband and wife be?

(a) 2

(b) 1

(c) 0.5

(d) 0

(e) Can't tell without seeing the data

31. The figure below is a scatterplot of reading test scores against IQ test scores for 14 fifth-grade children. There is one low outlier in the plot. What effect does this low outlier have on the correlation?

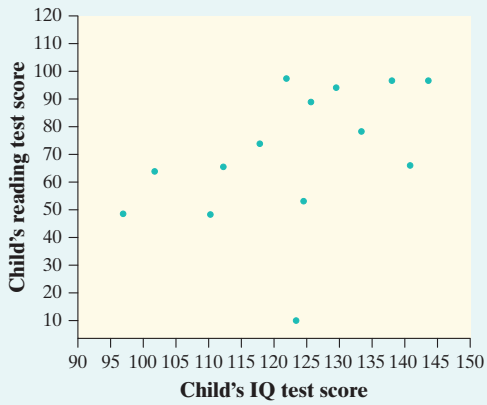
(a) It makes the correlation closer to 1.

(b) It makes the correlation closer to 0 but still positive.

(c) It makes the correlation equal to 0.

(d) It makes the correlation negative.

(e) It has no effect on the correlation.



32. If we leave out the low outlier, the correlation for the remaining 13 points in the preceding figure is closest to

(a) −0.95. (c) 0. (e) 0.95.

(b) −0.5. (d) 0.5.

33. **Big diamonds** (1.2, 1.3) Here are the weights (in milligrams) of 58 diamonds from a nodule carried up to the earth's surface in surrounding rock. These data represent a population of diamonds formed in a single event deep in the earth.[13]

13.8	3.7	33.8	11.8	27.0	18.9	19.3	20.8	25.4	23.1	7.8
10.9	9.0	9.0	14.4	6.5	7.3	5.6	18.5	1.1	11.2	7.0
7.6	9.0	9.5	7.7	7.6	3.2	6.5	5.4	7.2	7.8	3.5
5.4	5.1	5.3	3.8	2.1	2.1	4.7	3.7	3.8	4.9	2.4
1.4	0.1	4.7	1.5	2.0	0.1	0.1	1.6	3.5	3.7	2.6
4.0	2.3	4.5								

Make a graph that shows the distribution of weights of these diamonds. Describe what you see. Give appropriate numerical measures of center and spread.

34. College debt (2.2) A report published by the Federal Reserve Bank of New York in 2012 reported the results of a nationwide study of college student debt. Researchers found that the average student loan balance per borrower is $23,300. They also reported that about one-quarter of borrowers owe more than $28,000.[14]

(a) Assuming that the distribution of student loan balances is approximately Normal, estimate the standard deviation of the distribution of student loan balances.

(b) Assuming that the distribution of student loan balances is approximately Normal, use your answer to part (a) to estimate the proportion of borrowers who owe more than $54,000.

(c) In fact, the report states that about 10% of borrowers owe more than $54,000. What does this fact indicate about the shape of the distribution of student loan balances?

(d) The report also states that the median student loan balance is $12,800. Does this fact support your conclusion in part (c)? Explain.

3.2 Least-Squares Regression

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- Interpret the slope and y intercept of a least-squares regression line.
- Use the least-squares regression line to predict y for a given x. Explain the dangers of extrapolation.
- Calculate and interpret residuals.
- Explain the concept of least squares.
- Determine the equation of a least-squares regression line using technology or computer output.
- Construct and interpret residual plots to assess whether a linear model is appropriate.
- Interpret the standard deviation of the residuals and r^2 and use these values to assess how well the least-squares regression line models the relationship between two variables.
- Describe how the slope, y intercept, standard deviation of the residuals, and r^2 are influenced by outliers.
- Find the slope and y intercept of the least-squares regression line from the means and standard deviations of x and y and their correlation.

Linear (straight-line) relationships between two quantitative variables are fairly common and easy to understand. In the previous section, we found linear relationships in settings as varied as sparrowhawk colonies, natural-gas consumption, and Florida manatee deaths. Correlation measures the direction and strength of these relationships. When a scatterplot shows a linear relationship, we'd like to summarize the overall pattern by drawing a line on the scatterplot. A **regression line** summarizes the relationship between two variables, but only in a specific setting: when one of the variables helps explain or predict the other. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

DEFINITION: Regression line

A **regression line** is a line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x.

Let's look at a situation where a regression line provides a useful model.

EXAMPLE How Much Is That Truck Worth?

Regression lines as models

Everyone knows that cars and trucks lose value the more they are driven. Can we predict the price of a used Ford F-150 SuperCrew 4 × 4 if we know how many miles it has on the odometer? A random sample of 16 used Ford F-150 SuperCrew 4 × 4s was selected from among those listed for sale at autotrader.com. The number of miles driven and price (in dollars) were recorded for each of the trucks.[15] Here are the data:

Miles driven	70,583	129,484	29,932	29,953	24,495	75,678	8359	4447
Price (in dollars)	21,994	9500	29,875	41,995	41,995	28,986	31,891	37,991
Miles driven	34,077	58,023	44,447	68,474	144,162	140,776	29,397	131,385
Price (in dollars)	34,995	29,988	22,896	33,961	16,883	20,897	27,495	13,997

Figure 3.7 is a scatterplot of these data. The plot shows a moderately strong, negative linear association between miles driven and price with no outliers. The correlation is $r = -0.815$. The line on the plot is a regression line for predicting price from miles driven.

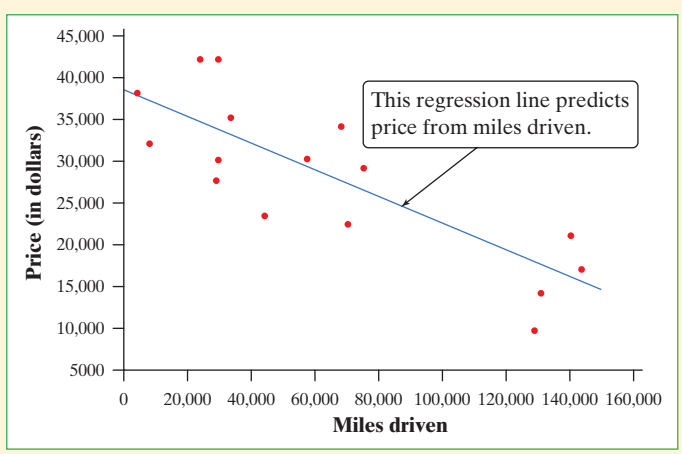


FIGURE 3.7 Scatterplot showing the price and miles driven of used Ford F-150s, with a regression line added.

Interpreting a Regression Line

A regression line is a *model* for the data, much like the density curves of Chapter 2. The equation of a regression line gives a compact mathematical description of what this model tells us about the relationship between the response variable y and the explanatory variable x.

DEFINITION: Regression line, predicted value, slope, *y* intercept

Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis). A **regression line** relating y to x has an equation of the form

$$\hat{y} = a + bx$$

In this equation,

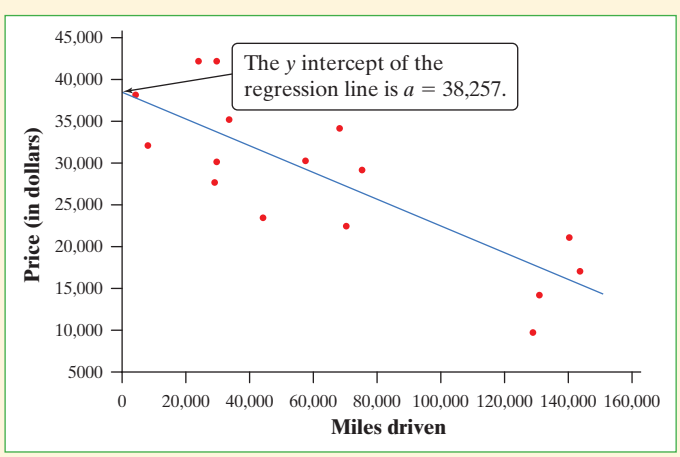
- $\hat{y}$ (read "y hat") is the **predicted value** of the response variable y for a given value of the explanatory variable x.
- b is the **slope**, the amount by which y is predicted to change when x increases by one unit.
- a is the ***y* intercept**, the predicted value of y when $x = 0$.

Although you are probably accustomed to the form $y = mx + b$ for the equation of a line from algebra, statisticians have adopted a different form for the equation of a regression line. Some use $\hat{y} = b_0 + b_1x$. We prefer $\hat{y} = a + bx$ for two reasons: (1) it's simpler and (2) your calculator uses this form. Don't get so caught up in the symbols that you lose sight of what they mean! The coefficient of x is always the slope, no matter what symbol is used.

Many calculators and software programs will give you the equation of a regression line from keyed-in data. Understanding and using the line are more important than the details of where the equation comes from.

EXAMPLE How Much Is That Truck Worth?

Interpreting the slope and y intercept



The equation of the regression line shown in Figure 3.7 is

$$\widehat{\text{price}} = 38{,}257 - 0.1629\ (\text{miles driven})$$

PROBLEM: Identify the slope and *y* intercept of the regression line. Interpret each value in context.

SOLUTION: The slope $b = -0.1629$ tells us that the price of a used Ford F-150 is *predicted* to go down by 0.1629 dollars (16.29 cents) for each additional mile that the truck has been driven. The *y* intercept $a = 38{,}257$ is the predicted price of a Ford F-150 that has been driven 0 miles.

For Practice *Try Exercise* 39(a) and (b)

The slope of a regression line is an important numerical description of the relationship between the two variables. Although we need the value of the y intercept to draw the line, it is statistically meaningful only when the explanatory variable can actually take values close to zero, as in this setting.

THINK ABOUT IT

Does a small slope mean that there's no relationship? For the miles driven and price regression line, the slope $b = -0.1629$ is a small number. This does *not* mean that change in miles driven has little effect on price. The size of the slope depends on the units in which we measure the two variables. In this setting, the slope is the predicted change in price (in dollars) when the distance driven increases by 1 mile. There are 100 cents in a dollar. If we measured price in cents instead of dollars, the slope would be 100 times larger, $b = 16.29$. *You can't say how strong a relationship is by looking at the size of the slope of the regression line.*

Prediction

We can use a regression line to predict the response $\hat{y}$ for a specific value of the explanatory variable x. Here's how we do it.

EXAMPLE

How Much Is That Truck Worth?

Predicting with a regression line

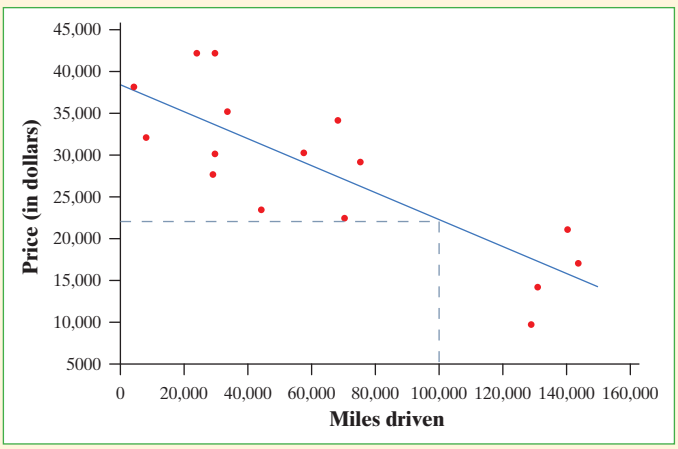


FIGURE 3.8 Using the regression line to predict price for a Ford F-150 with 100,000 miles driven.

For the Ford F-150 data, the equation of the regression line is

$$\widehat{\text{price}} = 38{,}257 - 0.1629\ (\text{miles driven})$$

If a used Ford F-150 has 100,000 miles driven, substitute $x = 100{,}000$ in the equation. The predicted price is

$$\widehat{\text{price}} = 38{,}257 - 0.1629(100{,}000) = 21{,}967 \text{ dollars}$$

This prediction is illustrated in Figure 3.8.

The accuracy of predictions from a regression line depends on how much the data scatter about the line. In this case, prices for trucks with similar mileage show a spread of about $10,000. The regression line summarizes the pattern but gives only roughly accurate predictions.

Can we predict the price of a Ford F-150 with 300,000 miles driven? We can certainly substitute 300,000 into the equation of the line. The prediction is

$$\widehat{\text{price}} = 38{,}257 - 0.1629(300{,}000) = -10{,}613 \text{ dollars}$$

That is, we predict that we would need to pay someone else $10,613 just to take the truck off our hands!

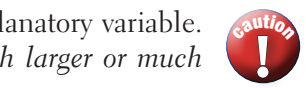
A negative price doesn't make much sense in this context. Look again at Figure 3.8. A truck with 300,000 miles driven is far outside the set of x values for our data. We can't say whether the relationship between miles driven and price remains linear at such extreme values. Predicting price for a truck with 300,000 miles driven is an **extrapolation** of the relationship beyond what the data show.

Often, using the regression line to make a prediction for $x = 0$ is an extrapolation. That's why the y intercept isn't always statistically meaningful.

DEFINITION: Extrapolation

Extrapolation is the use of a regression line for prediction far outside the interval of values of the explanatory variable x used to obtain the line. Such predictions are often not accurate.

Few relationships are linear for all values of the explanatory variable. *Don't make predictions using values of x that are much larger or much smaller than those that actually appear in your data.*

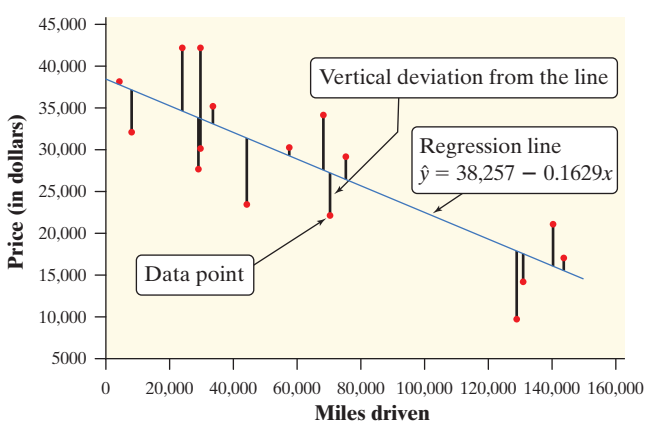


CHECK YOUR UNDERSTANDING

Some data were collected on the weight of a male white laboratory rat for the first 25 weeks after its birth. A scatterplot of the weight (in grams) and time since birth (in weeks) shows a fairly strong, positive linear relationship. The linear regression equation $\widehat{\text{weight}} = 100 + 40(\text{time})$ models the data fairly well.

1. What is the slope of the regression line? Explain what it means in context.
2. What's the y intercept? Explain what it means in context.
3. Predict the rat's weight after 16 weeks. Show your work.
4. Should you use this line to predict the rat's weight at age 2 years? Use the equation to make the prediction and think about the reasonableness of the result. (There are 454 grams in a pound.)

Residuals and the Least-Squares Regression Line



In most cases, no line will pass exactly through all the points in a scatterplot. Because we use the line to predict y from x, the prediction errors we make are errors in y, the vertical direction in the scatterplot. *A good regression line makes the vertical deviations of the points from the line as small as possible.*

Figure 3.9 shows a scatterplot of the Ford F-150 data with a regression line added. The prediction errors are

FIGURE 3.9 Scatterplot of the Ford F-150 data with a regression line added. A good regression line should make the prediction errors (shown as bold vertical segments) as small as possible.

marked as bold segments in the graph. These vertical deviations represent "left-over" variation in the response variable after fitting the regression line. For that reason, they are called **residuals.**

DEFINITION: Residual

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

$$\text{residual} = \text{observed } y - \text{predicted } y$$
$$= y - \hat{y}$$

The following example shows you how to calculate and interpret a residual.

EXAMPLE How Much Is That Truck Worth?

Finding a residual

PROBLEM: Find and interpret the residual for the Ford F-150 that had 70,583 miles driven and a price of $21,994.

SOLUTION: The regression line predicts a price of

$$\widehat{\text{price}} = 38{,}257 - 0.1629(70{,}583) = 26{,}759 \text{ dollars}$$

for this truck, but its actual price was $21,994. This truck's residual is

$$\text{residual} = \text{observed } y - \text{predicted } y$$
$$= y - \hat{y} = 21{,}994 - 26{,}759 = -4765 \text{ dollars}$$

That is, the actual price of this truck is $4765 lower than expected, based on its mileage. The actual price might be lower than predicted as a result of other factors. For example, the truck may have been in an accident or may need a new paint job.

For Practice *Try Exercise* 45

The line shown in Figure 3.9 makes the residuals for the 16 trucks "as small as possible." But what does that mean? Maybe this line minimizes the *sum* of the residuals. Actually, if we add up the prediction errors for all 16 trucks, the positive and negative residuals cancel out. That's the same issue we faced when we tried to measure deviation around the mean in Chapter 1. We'll solve the current problem in much the same way: by squaring the residuals. The regression line we want is the one that minimizes the sum of the squared residuals. That's what the line shown in Figure 3.9 does for the Ford F-150 data, which is why we call it the **least-squares regression line**.

DEFINITION: Least-squares regression line

The **least-squares regression line** of y on x is the line that makes the sum of the squared residuals as small as possible.

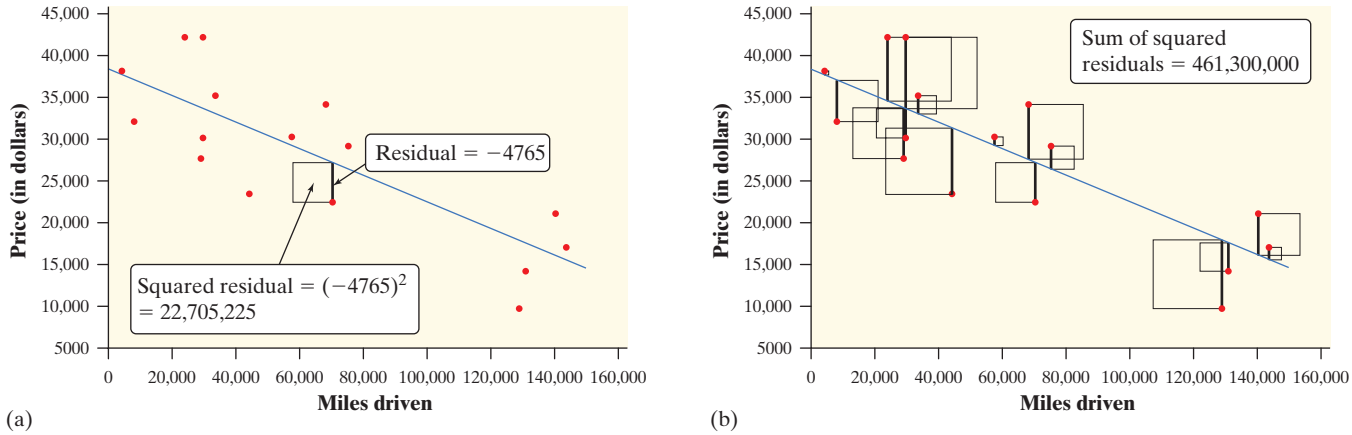


FIGURE 3.10 The least-squares idea: make the errors in predicting y as small as possible by minimizing the sum of the squares of the residuals.

Figure 3.10 gives a geometric interpretation of the least-squares idea for the truck data. Figure 3.10(a) shows the "squared" residual for the truck with 70,583 miles driven and a price of $21,994. The area of this square is $(-4765)(-4765) = 22{,}705{,}225$. Figure 3.10(b) shows the squared residuals for all the trucks. The sum of squared residuals is 461,300,000. No other regression line would give a smaller sum of squared residuals.

ACTIVITY Investigating properties of the least-squares regression line

MATERIALS:
Computer with
Internet connection

In this Activity, you will use the *Correlation and Regression* applet at the book's Web site, www.whfreeman.com/tps5e, to explore some properties of the least-squares regression line.

1. Click on the scatterplot to create a group of 15 to 20 points from lower left to upper right with a clear positive straight-line pattern (correlation around 0.7).

2. Click the "Draw your own line" button to select starting and ending points for your own line on the plot. Use the mouse to adjust the starting and ending points until you have a line that models the association well.

3. Click the "Show least-squares line" button. How do the two lines compare? One way to measure this is to compare the "Relative SS," the ratio of the sum of squared residuals from your line and the least-squares regression line. If the two lines are exactly the same, the relative sum of squares will be 1. Otherwise, the relative sum of squares will be larger than 1.

4. Press the "CLEAR" button and create another scatterplot as in Step 1. Then click on "Show least-squares line" and "Show mean X & Y lines." What do you notice? Move or add points, one at a time, in your scatterplot to see if this result continues to hold true.

5. Now click the "Show residuals" button. How does an outlier affect the slope and y intercept of the least-squares regression line? Move or add points, one at a time, to investigate. Does it depend on whether the outlier has an x-value close to the center of the plot or toward the far edges of the plot?

Your calculator or statistical software will give the equation of the least-squares line from data that you enter. Then you can concentrate on understanding and using the regression line.

8. TECHNOLOGY CORNER LEAST-SQUARES REGRESSION LINES ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

Let's use the Ford F-150 data to show how to find the equation of the least-squares regression line on the TI-83/84 and TI-89. Here are the data again:

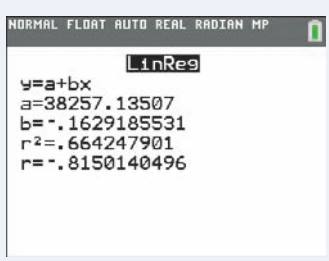
Miles driven	70,583	129,484	29,932	29,953	24,495	75,678	8359	4447
Price (in dollars)	21,994	9500	29,875	41,995	41,995	28,986	31,891	37,991
Miles driven	34,077	58,023	44,447	68,474	144,162	140,776	29,397	131,385
Price (in dollars)	34,995	29,988	22,896	33,961	16,883	20,897	27,495	13,997

1. Enter the miles driven data into L1/list1 and the price data into L2/list2. Then make a scatterplot. Refer to the Technology Corner on page 150.

2. To determine the least-squares regression line:

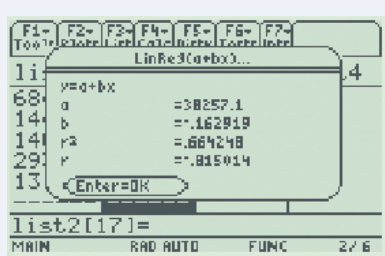
TI-83/84

- Press STAT; choose CALC and then `LinReg(a+bx)`. **OS 2.55 or later:** In the dialog box, enter the following: `Xlist:L1`, `Ylist:L2`, `FreqList (leave blank)`, `Store RegEQ:Y1`, and choose `Calculate`. **Older OS:** Finish the command to read `LinReg(a+bx) L1,L2,Y1` and press ENTER. (Y1 is found under VARS/Y-VARS/Function.)



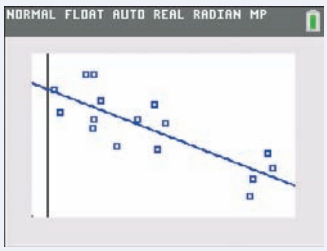
TI-89

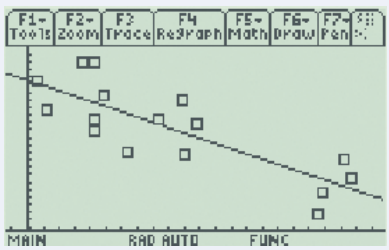
- In the Statistics/List Editor, press F4 (CALC); choose `Regressions` and then `LinReg(a+bx)`.
- Enter list1 for the Xlist, list2 for the Ylist; choose to store the RegEqn to y1(x); and press ENTER



Note: If you do not want to store the equation to Y1, then leave the StoreRegEq prompt blank (OS 2.55 or later) or use the following command (older OS): `LinReg(a+bx) L1,L2`.

3. Graph the regression line. Turn off all other equations in the Y= screen and use ZoomStat/ZoomData to add the least-squares line to the scatterplot.





4. Save these lists for later use. On the home screen, use the STO▸ key to help execute the command `L1→MILES:L2→PRICE` (`list1→MILES:list2→PRICE` on the TI-89).

Note: If r^2 and r do not appear on the TI-83/84 screen, do this one-time series of keystrokes: **OS 2.55 or later:** Press MODE and set STAT DIAGNOSTICS to ON. **Older OS:** Press 2nd 0 (CATALOG), scroll down to `DiagnosticOn`, and press ENTER. Press ENTER again to execute the command. The screen should say "Done." Then redo Step 2 to calculate the least-squares line. The r^2 and r values should now appear.

AP® EXAM TIP When displaying the equation of a least-squares regression line, the calculator will report the slope and intercept with much more precision than we need. However, there is no firm rule for how many decimal places to show for answers on the AP® exam. Our advice: Decide how much to round based on the context of the problem you are working on.

CHECK YOUR UNDERSTANDING

It's time to practice your calculator regression skills. Using the familiar SEC football data in the table below, repeat the steps in the previous Technology Corner. You should get $\hat{y} = -3.7506 + 0.4372x$ as the equation of the regression line.

Team	Alabama	Arkansas	Auburn	Florida	Georgia	Kentucky
Points per game	34.8	36.8	25.7	25.5	32.0	15.8
Wins	12	11	8	7	10	5

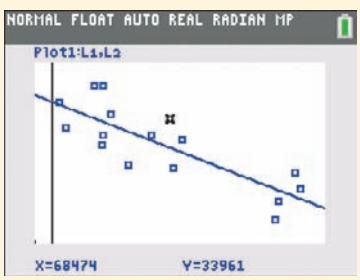
Team	Louisiana State	Mississippi	Mississippi State	South Carolina	Tennessee	Vanderbilt
Points per game	35.7	16.1	25.3	30.1	20.3	26.7
Wins	13	2	7	11	5	6

Determining Whether a Linear Model Is Appropriate: Residual Plots

One of the first principles of data analysis is to look for an overall pattern and for striking departures from the pattern. A regression line describes the overall pattern of a linear relationship between an explanatory variable and a response variable. We see departures from this pattern by looking at the residuals.

EXAMPLE How Much Is That Truck Worth?

Examining residuals



Let's return to the Ford F-150 data about the number of miles driven and price for a random sample of 16 used trucks. In general, trucks with more miles driven have lower prices. In the Technology Corner, we confirmed that the equation of the least-squares regression line for these data is $\widehat{\text{price}} = 38{,}257 - 0.1629$ (miles driven). The calculator screen shot in the margin shows a scatterplot of the data with the least-squares line added.

One truck had 68,474 miles driven and a price of $33,961. This truck is marked on the scatterplot with an X. Because the point is above the line on the scatterplot, we know that its actual price is higher than the predicted price. To find out exactly how much higher, we calculate the residual for this truck. The predicted price for a Ford F-150 with 68,474 miles driven is

$$\hat{y} = 38{,}257 - 0.1629(68{,}474) = \$27{,}103$$

The residual for this truck is therefore

$$\text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y} = 33{,}961 - 27{,}103 = \$6858$$

This truck costs $6858 more than expected, based on its mileage.

The 16 points used in calculating the equation of the least-squares regression line produce 16 residuals. Rounded to the nearest dollar, they are

−4765	−7664	−3506	8617	7728	3057	−5004	458
2289	1183	−8121	6858	2110	5572	−5973	−2857

Most graphing calculators and statistical software will calculate and store residuals for you.

Although residuals can be calculated from any model that is fitted to the data, the residuals from the least-squares line have a special property: *the mean of the least-squares residuals is always zero*. You can check that the sum of the residuals in the above example is −$18. The sum is not exactly 0 because of rounding errors.

You can see the residuals in the scatterplot of Figure 3.11(a) on the next page by looking at the vertical deviations of the points from the line. The **residual plot** in Figure 3.11(b) makes it easier to study the residuals by plotting them against the explanatory variable, miles driven. Because the mean of the residuals is always zero, the horizontal line at zero in Figure 3.11(b) helps orient us. This "residual = 0" line corresponds to the regression line in Figure 3.11(a).

Some software packages prefer to plot the residuals against the predicted values $\hat{y}$ instead of against the values of the explanatory variable. The basic shape of the two plots is the same because $\hat{y}$ is linearly related to x.

DEFINITION: Residual plot

A **residual plot** is a scatterplot of the residuals against the explanatory variable. Residual plots help us assess whether a linear model is appropriate.

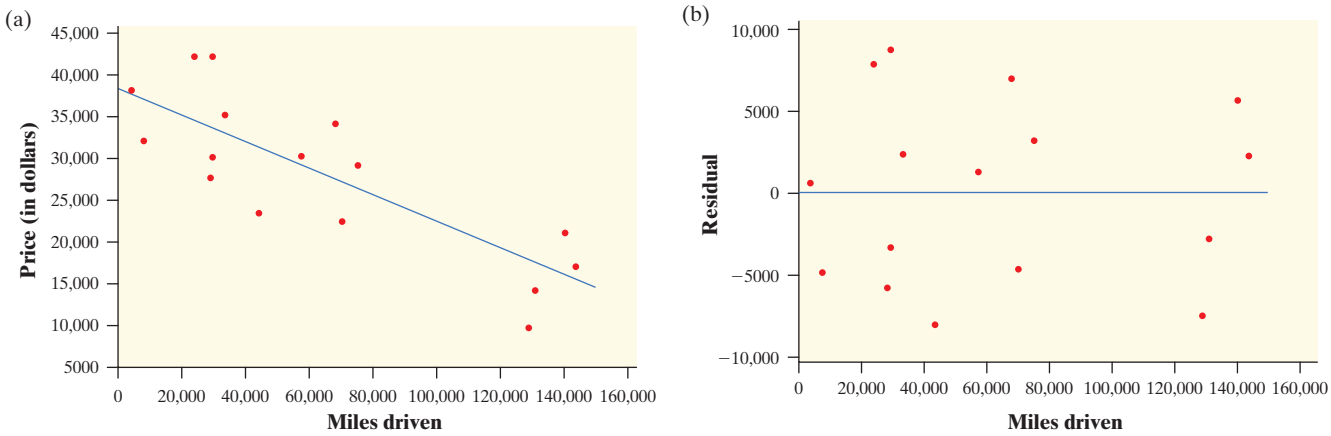


FIGURE 3.11 (a) Scatterplot of price versus miles driven, with the least-squares line. (b) Residual plot for the regression line displayed in Figure 3.11(a). The line at $y = 0$ marks the sum (and mean) of the residuals.

CHECK YOUR UNDERSTANDING

Refer to the Ford F-150 miles driven and price data.

1. Find the residual for the truck that had 8359 miles driven and a price of $31,891. Show your work.
2. Interpret the value of this truck's residual in context.
3. For which truck did the regression line overpredict price by the most? Justify your answer.

Examining residual plots

A residual plot in effect turns the regression line horizontal. It magnifies the deviations of the points from the line, making it easier to see unusual observations and patterns. Because it is easier to see an unusual pattern in a residual plot than a scatterplot of the original data, we often use residual plots to determine if the model we are using is appropriate.

Figure 3.12(a) shows a nonlinear association between two variables and the least-squares regression line for these data. Figure 3.12(b) shows the residual plot for these data.

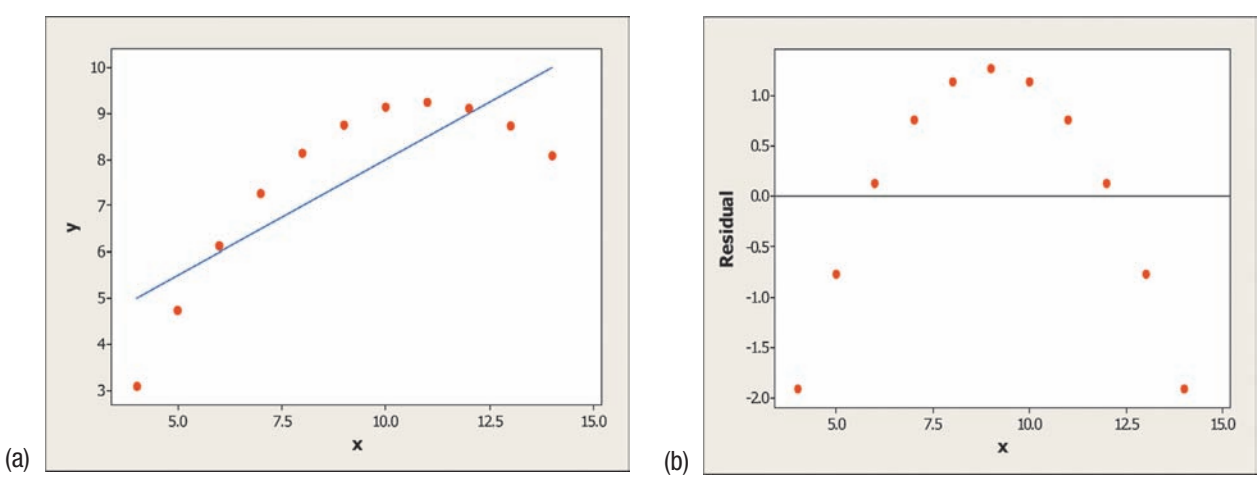


FIGURE 3.12 (a) A straight line is not a good model for these data. (b) The residual plot has a curved pattern.

Because the form of our model (linear) is not the same as the form of the association (curved), there is an obvious leftover pattern in the residual plot. *When an obvious curved pattern exists in a residual plot, the model we are using is not appropriate.* We'll look at how to deal with curved relationships in Chapter 12.

When we use a line to model a linear association, there will be no leftover patterns in the residual plot, only random scatter. Figure 3.13 shows the residual plot for the Ford F-150 data. Because there is only random scatter in the residual plot, we know the linear model we used is appropriate.

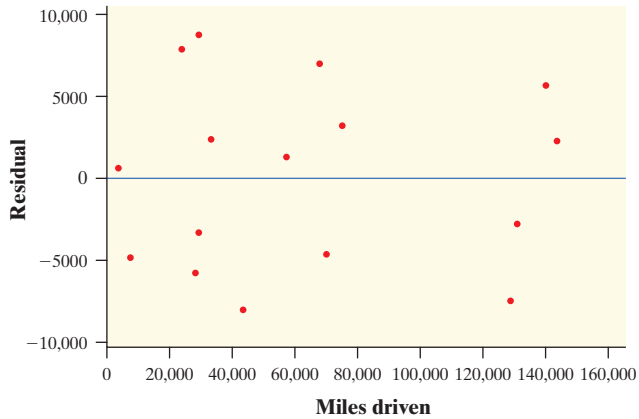


FIGURE 3.13 The random scatter of points indicates that the regression line has the same form as the association, so the line is an appropriate model.

THINK ABOUT IT

Why do we look for patterns in residual plots? The word *residual* comes from the Latin word *residuum*, meaning "left over." When we calculate a residual, we are calculating what is left over after subtracting the predicted value from the observed value:

$$\text{residual} = \text{observed } y - \text{predicted } y$$

Likewise, when we look at the form of a residual plot, we are looking at the form that is left over after subtracting the form of the model from the form of the association:

$$\text{form of residual plot} = \text{form of association} - \text{form of model}$$

When there is a leftover form in the residual plot, the form of the association and form of the model are not the same. However, if the form of the association and form of the model are the *same*, the residual plot should have no form, other than random scatter.

9. TECHNOLOGY CORNER RESIDUAL PLOTS ON THE CALCULATOR

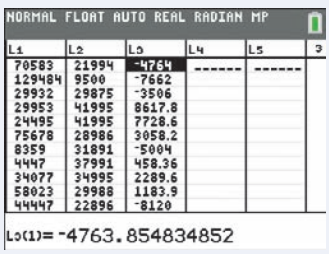
TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

Let's continue the analysis of the Ford F-150 miles driven and price data from the previous Technology Corner (page 171). You should have already made a scatterplot, calculated the equation of the least-squares regression line, and graphed the line on your plot. Now, we want to calculate residuals and make a residual plot. Fortunately, your calculator has already done most of the work. Each time the calculator computes a regression line, it also computes the residuals and stores them in a list named RESID. Make sure to calculate the equation of the regression line *before* using the RESID list!

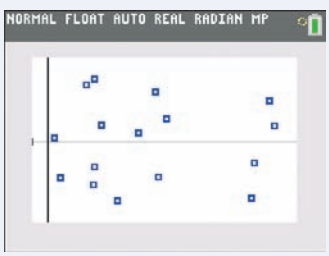
TI-83/84

1. Display the residuals in L3(list3).

- With L3 highlighted, press 2nd STAT (LIST) and select the RESID list.



2. Turn off Plot1 and the regression equation. Specify Plot2 with L1/list1 as the *x* variable and L3/list3 as the *y* variable. Use ZoomStat (ZoomData) to see the residual plot.

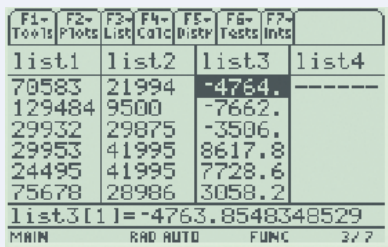


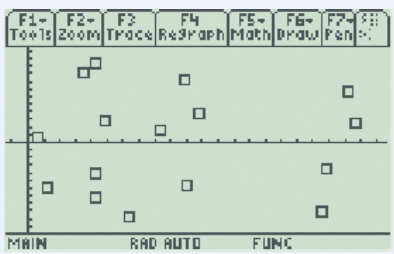
The *x* axis in the residual plot serves as a reference line: points above this line correspond to positive residuals and points below the line correspond to negative residuals.

Note: If you don't want to see the residuals in L3/list3, you can make a residual plot in one step by using the RESID list as the *y* variable in the scatterplot.

TI-89

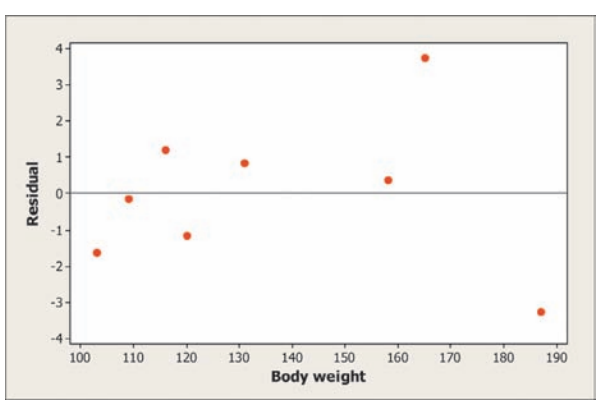
- With list3 highlighted, press 2nd - (VAR-LINK), arrow down to STATVARS, and select the RESID list.





CHECK YOUR UNDERSTANDING

In Exercises 5 and 7, we asked you to make and describe a scatterplot for the hiker data shown in the table below. Here is a residual plot for the least-squares regression of pack weight on body weight for the 8 hikers.



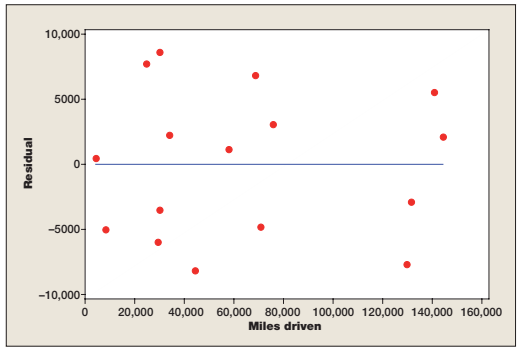
Body weight (lb):	120	187	109	103	131	165	158	116
Backpack weight (lb):	26	30	26	24	29	35	31	28

1. One of the hikers had a residual of nearly 4 pounds. Interpret this value.
2. Based on the residual plot, is a linear model appropriate for these data?

How Well the Line Fits the Data: The Role of s and r^2 in Regression

A residual plot is a graphical tool for determining if a least-squares regression line is an appropriate model for a relationship between two variables. Once we determine that a least-squares regression line is appropriate, it makes sense to ask a follow-up question: How well does the line work? That is, if we use the least-squares regression line to make predictions, how good will these predictions be?

The Standard Deviation of the Residuals We already know that a residual measures how far an observed y-value is from its corresponding predicted value $\hat{y}$. In an earlier example, we calculated the residual for the Ford F-150 with 68,474 miles driven and price \$33,961. The residual was \$6858, meaning that the actual price was \$6858 higher than we predicted.



To assess how well the line fits *all* the data, we need to consider the residuals for each of the 16 trucks, not just one. Using these residuals, we can estimate the "typical" prediction error when using the least-squares regression line. To do this, we calculate the **standard deviation of the residuals.**

$$s = \sqrt{\frac{\Sigma \text{residuals}^2}{n-2}}$$

For the Ford F-150 data, the sum of squared residuals is 461,300,000. So, the standard deviation of the residuals is

$$s = \sqrt{\frac{461{,}300{,}000}{14}} = 5740 \text{ dollars}$$

Did you recognize the number 461,300,000? We first encountered this number on page 170 when illustrating that the least-squares regression line minimized the sum of squared residuals. We'll see it again shortly.

When we use the least-squares regression line to predict the price of a Ford F-150 using the number of miles it has been driven, our predictions will typically be off by about \$5740. Looking at the residual plot, this seems like a reasonable value. Although some of the residuals are close to 0, others are close to \$10,000 or −\$10,000.

DEFINITION: Standard deviation of the residuals (s)

If we use a least-squares line to predict the values of a response variable y from an explanatory variable x, the **standard deviation of the residuals (s)** is given by

$$s = \sqrt{\frac{\Sigma \text{ residuals}^2}{n-2}} = \sqrt{\frac{\Sigma (y_i - \hat{y})^2}{n-2}}$$

This value gives the approximate size of a "typical" prediction error (residual).

THINK ABOUT IT

Does the formula for s look slightly familiar? It should. In Chapter 1, we defined the standard deviation of a set of quantitative data as

$$s_x = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$$

We interpreted the resulting value as the "typical" distance of the data points from the mean. In the case of two-variable data, we're interested in the typical (vertical) distance of the data points from the regression line. We find this value in much the same way: by adding up the squared deviations, then averaging (again in a funny way), and taking the square root to get back to the original units of measurement. Why do we divide by $n - 2$ this time instead of $n - 1$? You'll have to wait until Chapter 12 to find out.

The Coefficient of Determination There is another numerical quantity that tells us how well the least-squares line predicts values of the response variable y. It is r^2, the coefficient of determination. Some computer packages call it "R-sq." You may have noticed this value in some of the calculator and computer regression output that we showed earlier. Although it's true that r^2 is equal to the square of r, there is much more to this story.

EXAMPLE How Much Is That Truck Worth?

How can we predict y if we don't know x?

Suppose that we randomly selected an additional used Ford F-150 that was on sale. What should we predict for its price? Figure 3.14 shows a scatterplot of the truck data that we have studied throughout this section, including the least-squares regression line. Another horizontal line has been added at the mean y-value, $\bar{y} = \$27{,}834$. If we don't know the number of miles driven for the additional truck, we can't use the regression line to make a prediction. What should we do? Our best strategy is to use the mean price of the other 16 trucks as our prediction.

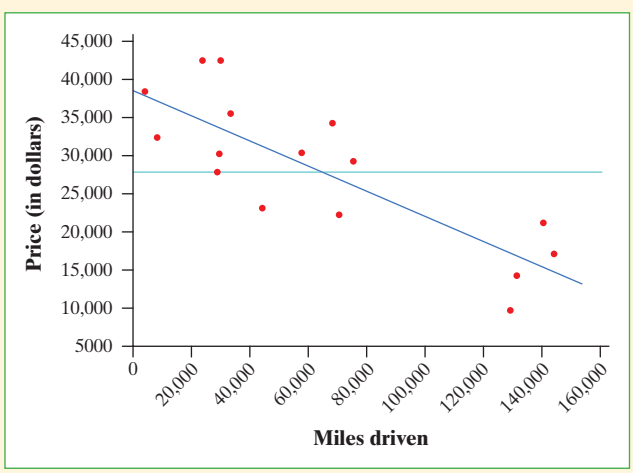


FIGURE 3.14 Scatterplot and least-squares regression line for the Ford F-150 data with a horizontal line added at the mean price, $27,834.

Figure 3.15(a) on the facing page shows the prediction errors if we use the average price $\bar{y}$ as our prediction for the original group of 16 trucks. We can see that the sum of the squared residuals for this line is $\Sigma(y_i - \bar{y})^2 = 1{,}374{,}000{,}000$. This quantity measures the total variation in the y-values from their mean. This is also the same quantity we use to calculate the standard deviation of the prices, s_y.

If we learn the number of miles driven on the additional truck, then we could use the least-squares line to predict its price. How much better does the regression line do at predicting prices than simply using the average price $\bar{y}$ of all 16 trucks? Figure 3.15(b) reminds us that the sum of squared residuals for the least-squares line is $\Sigma\text{residuals}^2 = 461{,}300{,}000$. This is the same quantity we used to calculate the standard deviation of the residuals.

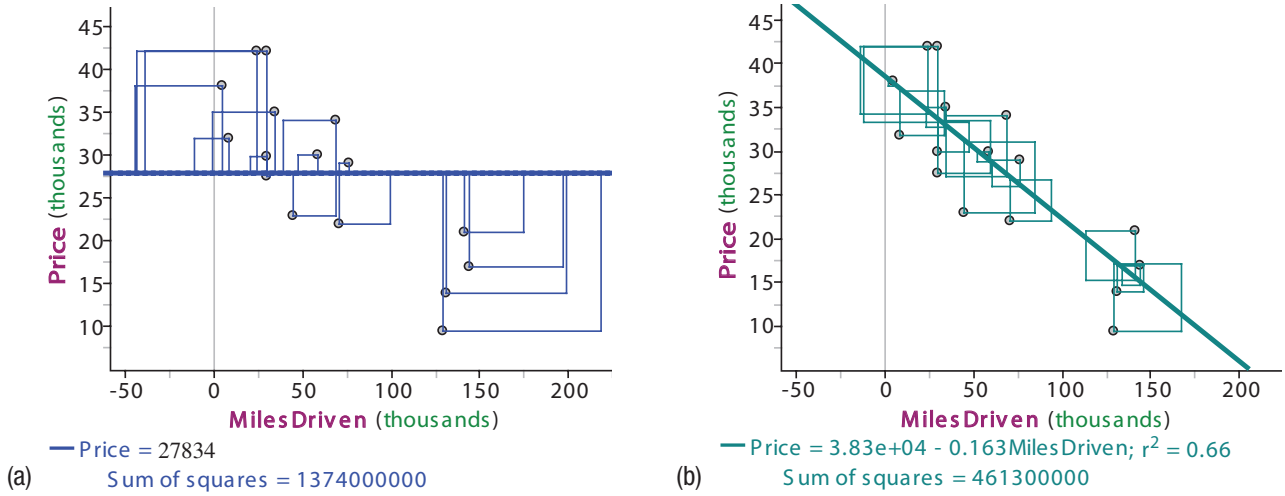


FIGURE 3.15 (a) The sum of squared residuals is 1,374,000,000 if we use the mean price as our prediction for all 16 trucks. (b) The sum of squares from the least-squares regression line is 461,300,000.

The ratio of these two quantities tells us what proportion of the total variation in y still remains after using the regression line to predict the values of the response variable. In this case,

$$\frac{461{,}300{,}000}{1{,}374{,}000{,}000} = 0.336$$

This means that 33.6% of the variation in price is *unaccounted for* by the least-squares regression line using x = miles driven. This unaccounted-for variation is likely due to other factors, including the age of the truck or its condition. Taking this one step further, the proportion of the total variation in y that *is accounted for* by the regression line is

$$1 - 0.336 = 0.664$$

We interpret this by saying that "*66.4% of the variation in price is accounted for by the linear model relating price to miles driven.*"

DEFINITION: The coefficient of determination: r^2

The **coefficient of determination** r^2 is the fraction of the variation in the values of y that is accounted for by the least-squares regression line of y on x. We can calculate r^2 using the following formula:

$$r^2 = 1 - \frac{\sum \text{residuals}^2}{\sum (y_i - \bar{y})^2}$$

If all the points fall directly on the least-squares line, the sum of squared residuals is 0 and $r^2 = 1$. Then all the variation in y is accounted for by the linear relationship with x. Because the least-squares line yields the smallest possible sum of squared prediction errors, the sum of squared residuals can never be more than the sum of squared deviations from the mean of y. In the worst-case scenario, the least-squares line does no better at predicting y than $y = \bar{y}$ does. Then the two sums of squares are the same and $r^2 = 0$.

It seems fairly remarkable that the coefficient of determination is actually the correlation squared. This fact provides an important connection between correlation and regression. When you see a correlation, square it to get a better feel for how well the least-squares line fits the data.

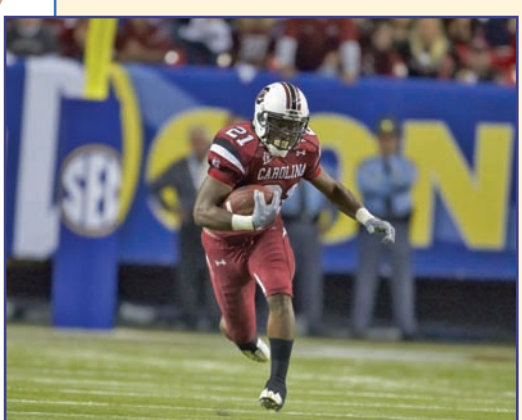
THINK ABOUT IT

What's the relationship between the standard deviation of the residuals s and the coefficient of determination r^2? They are both calculated from the sum of squared residuals. They also both attempt to answer the question, "How well does the line fit the data?" The standard deviation of the residuals reports the size of a typical prediction error, in the same units as the response variable. In the truck example, $s = 5740$ *dollars*. The value of r^2, however, does not have units and is usually expressed as a percentage between 0% and 100%, such as $r^2 = 66.4\%$. Because these values assess how well the line fits the data in different ways, we recommend you follow the example of most statistical software and report them both.

Let's revisit the SEC football data to practice what we have learned.

EXAMPLE SEC Football

Residual plots, s, and r^2



In Section 3.1, we looked at the relationship between the average number of points scored per game x and the number of wins y for the 12 college football teams in the Southeastern Conference. A scatterplot with the least-squares regression line and a residual plot are shown. The equation of the least-squares regression line is $\hat{y} = -3.75 + 0.437x$. Also, $s = 1.24$ and $r^2 = 0.88$.

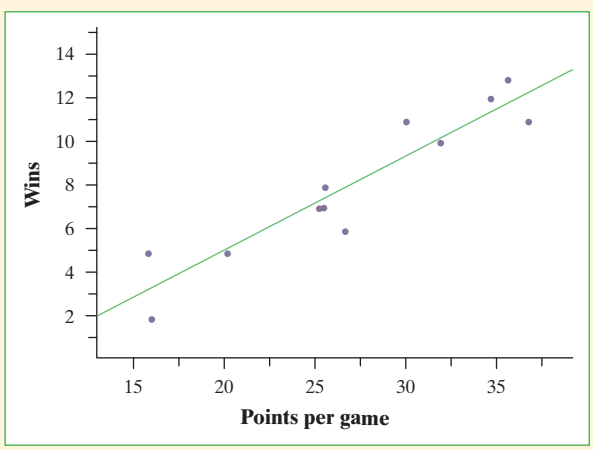
PROBLEM:

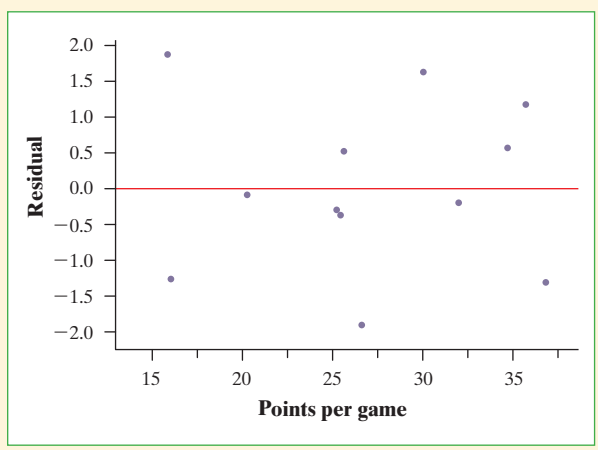
(a) Calculate and interpret the residual for South Carolina, which scored 30.1 points per game and had 11 wins.

(b) Is a linear model appropriate for these data? Explain.

(c) Interpret the value of s.

(d) Interpret the value of r^2.





AP® EXAM TIP Students often have a hard time interpreting the value of r^2 on AP® exam questions. They frequently leave out key words in the definition. Our advice: Treat this as a fill-in-the-blank exercise. Write "____% of the variation in [response variable name] is accounted for by the linear model relating [response variable name] to [explanatory variable name]."

SOLUTION:

(a) The predicted amount of wins for South Carolina is

$$\hat{y} = -3.75 + 0.437(30.1) = 9.40 \text{ wins}$$

The residual for South Carolina is

$$\text{residual} = y - \hat{y} = 11 - 9.40 = 1.60 \text{ wins}$$

South Carolina won 1.60 more games than expected, based on the number of points they scored per game.

(b) Because there is no obvious pattern left over in the residual plot, the linear model is appropriate.

(c) When using the least-squares regression line with x = points per game to predict y = the number of wins, we will typically be off by about 1.24 wins.

(d) About 88% of the variation in wins is accounted for by the linear model relating wins to points per game.

For Practice *Try Exercise* 55

Interpreting Computer Regression Output

Figure 3.16 displays the basic regression output for the Ford F-150 data from two statistical software packages: Minitab and JMP. Other software produces very similar output. Each output records the slope and y intercept of the least-squares line. The software also provides information that we don't yet need (or understand!), although we will use much of it later. Be sure that you can locate the slope, the y intercept, and the values of s and r^2 on both computer outputs. *Once you understand the statistical ideas, you can read and work with almost any software output.*

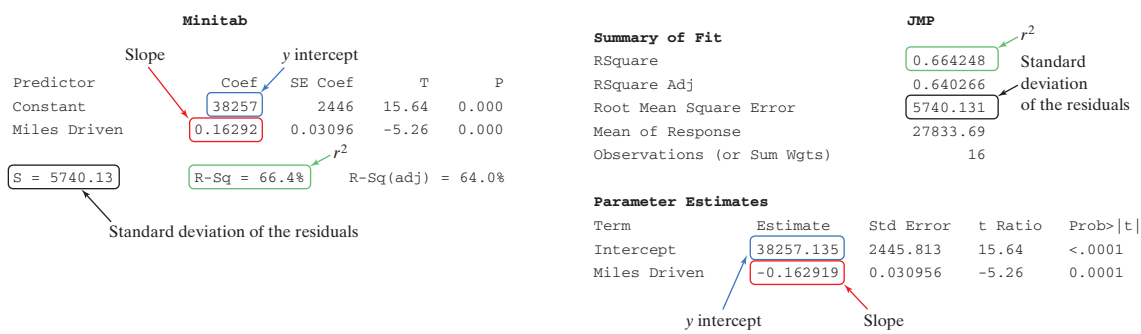


FIGURE 3.16 Least-squares regression results for the Ford F-150 data from two statistical software packages. Other software produces similar output.

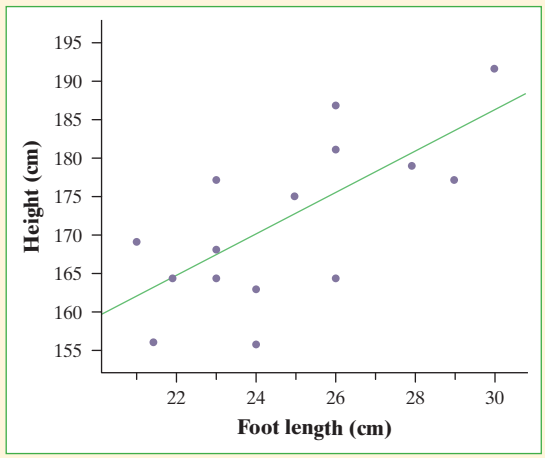
EXAMPLE Using Feet to Predict Height

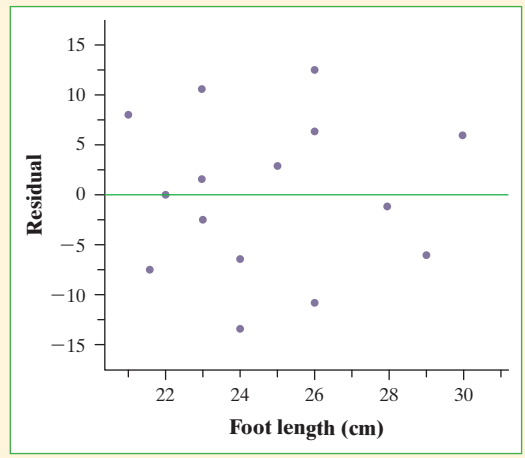
Interpreting regression output

A random sample of 15 high school students was selected from the U.S. CensusAtSchool database. The foot length (in centimeters) and height (in centimeters) of each student in the sample were recorded. Least-squares

regression was performed on the data. A scatterplot with the regression line added, a residual plot, and some computer output from the regression are shown below.

```
Predictor          Coef       SE Coef     T          P
Constant           103.41     19.50       5.30       0.000
Foot length        2.7469     0.7833      3.51       0.004
S = 7.95126     R-Sq = 48.6%     R-Sq(adj) = 44.7%
```





PROBLEM:

(a) What is the equation of the least-squares regression line that describes the relationship between foot length and height? Define any variables that you use.

(b) Interpret the slope of the regression line in context.

(c) Find the correlation.

(d) Is a line an appropriate model to use for these data? Explain how you know.

SOLUTION:

(a) The equation is $\hat{y} = 103.41 + 2.7469x$, where $\hat{y}$ = predicted height (in centimeters) and x is foot length (in centimeters). We could also write

$$\text{predicted height} = 103.41 + 2.7469\,(\text{foot length})$$

(b) For each additional centimeter of foot length, the least-squares regression line predicts an increase of 2.7469 cm in height.

(c) To find the correlation, we take the square root of r^2: $r = \pm\sqrt{0.486} = \pm 0.697$. Because the scatterplot shows a positive association, $r = 0.697$.

(d) Because the scatterplot shows a linear association and the residual plot has no obvious leftover patterns, a line is an appropriate model to use for these data.

For Practice *Try Exercise* 59

Regression to the Mean

Using technology is often the most convenient way to find the equation of a least-squares regression line. It is also possible to calculate the equation of the least-squares regression line using only the means and standard deviations of the two

variables and their correlation. Exploring this method will highlight an important relationship between the correlation and the slope of a least-squares regression line—and reveal why we include the word "regression" in the expression "least-squares regression line."

HOW TO CALCULATE THE LEAST-SQUARES REGRESSION LINE

We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations s_x and s_y of the two variables and their correlation r. The least-squares regression line is the line $\hat{y} = a + bx$ with **slope**

$$b = r\frac{s_y}{s_x}$$

and **y intercept**

$$a = \bar{y} - b\bar{x}$$

AP® EXAM TIP The formula sheet for the AP® exam uses different notation for these equations: $b_1 = r\frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$. That's because the least-squares line is written as $\hat{y} = b_0 + b_1x$. We prefer our simpler versions without the subscripts!

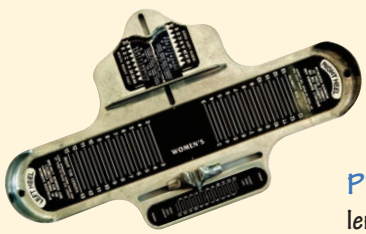
The formula for the y intercept comes from the fact that the least-squares regression line always passes through the point $(\bar{x}, \bar{y})$. You discovered this in Step 4 of the Activity on page 170. Substituting $(\bar{x}, \bar{y})$ into the equation $\hat{y} = a + bx$ produces the equation $\bar{y} = a + b\bar{x}$. Solving this equation for a gives the equation shown in the definition box, $a = \bar{y} - b\bar{x}$.

To see how these formulas work in practice, let's look at an example.

EXAMPLE Using Feet to Predict Height

Calculating the least-squares regression line

In the previous example, we used data from a random sample of 15 high school students to investigate the relationship between foot length (in centimeters) and height (in centimeters). The mean and standard deviation of the foot lengths are $\bar{x} = 24.76$ cm and $s_x = 2.71$ cm. The mean and standard deviation of the heights are $\bar{y} = 171.43$ cm and $s_y = 10.69$ cm. The correlation between foot length and height is $r = 0.697$.



PROBLEM: Find the equation of the least-squares regression line for predicting height from foot length. Show your work.

SOLUTION: The least-squares regression line of height y on foot length x has slope

$$b = r\frac{s_y}{s_x} = 0.697\frac{10.69}{2.71} = 2.75$$

The least-squares regression line has y intercept

$$a = \bar{y} - b\bar{x} = 171.43 - 2.75(24.76) = 103.34$$

So, the equation of the least-squares regression line is $\hat{y} = 103.34 + 2.75x$.

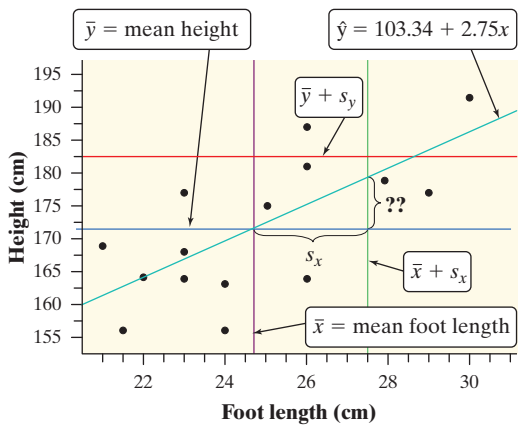
For Practice *Try Exercise* 61(a)

There is a close connection between the correlation and the slope of the least-squares regression line. The slope is

$$b = r\frac{s_y}{s_x} = \frac{r \cdot s_y}{s_x}$$

This equation says that along the regression line, a change of 1 standard deviation in x corresponds to a change of r standard deviations in y. When the variables are perfectly correlated ($r = 1$ or $r = -1$), the change in the predicted response $\hat{y}$ is the same (in standard deviation units) as the change in x. For example, if $r = 1$ and x is 2 standard deviations above its mean, then the corresponding value of $\hat{y}$ will be 2 standard deviations above the mean of y.

However, if the variables are not perfectly correlated ($-1 < r < 1$), the change in $\hat{y}$ is *less than* the change in x, when measured in standard deviation units. To illustrate this property, let's return to the foot length and height data from the previous example.



The figure at left shows the regression line $\hat{y} = 103.34 + 2.75x$. We have added four more lines to the graph: a vertical line at the mean foot length $\bar{x}$, a vertical line at $\bar{x} + s_x$ (1 standard deviation above the mean foot length), a horizontal line at the mean height $\bar{y}$, and a horizontal line at $\bar{y} + s_y$ (1 standard deviation above the mean height).

When a student's foot length is 1 standard deviation above the mean foot length $\bar{x}$, the predicted height $\hat{y}$ is above the mean height $\bar{y}$, but not an entire standard deviation above the mean. How far above the mean is the value of $\hat{y}$?

From the graph, we can see that

$$b = \text{slope} = \frac{\text{change in } y}{\text{change in } x} = \frac{??}{s_x}$$

From earlier, we know that

$$b = \frac{r \cdot s_y}{s_x}$$

Setting these two equations equal to each other, we have

$$\frac{??}{s_x} = \frac{r \cdot s_y}{s_x}$$

Thus, $\hat{y}$ must be $r \cdot s_y$ above the mean $\bar{y}$.

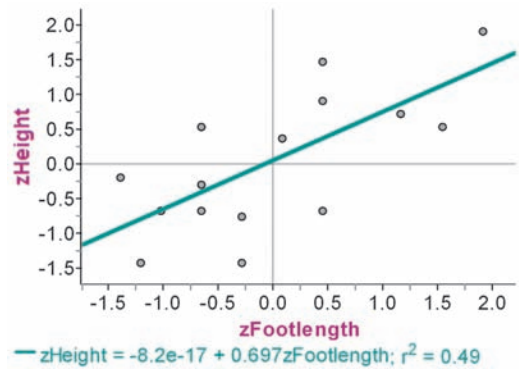
In other words, for an increase of 1 standard deviation in the value of the explanatory variable x, the least-squares regression line predicts an increase of *only* r standard deviations in the response variable y. *When the correlation isn't $r = 1$ or -1, the predicted value of y is closer to its mean $\bar{y}$ than the value of x is to its mean $\bar{x}$. This is called regression to the mean, because the values of y "regress" to their mean.*

Sir Francis Galton (1822–1911) looked at data on the heights of children versus the heights of their parents. He found that taller-than-average parents tended to have children who were taller than average but not quite as tall as their parents. Likewise, shorter-than-average parents tended to have children who were shorter than average but not quite as short as their parents. Galton called this fact "regression to the mean" and used the symbol r because of the correlation's important relationship to regression.

THINK ABOUT IT

What happens if we standardize both variables? Standardizing a variable converts its mean to 0 and its standard deviation to 1. Doing this to both x and y will transform the point $(\bar{x}, \bar{y})$ to $(0, 0)$. So the least-squares line for the standardized values will pass through $(0, 0)$. What about the slope of this line? From the formula, it's $b = rs_y/s_x$. Because we standardized, $s_x = s_y = 1$. That

means $b = r$. In other words, the slope is equal to the correlation. The Fathom screen shot confirms these results. It shows that $r^2 = 0.49$, so $r = \sqrt{0.49} = 0.7$, approximately the same value as the slope of 0.697.



Putting It All Together: Correlation and Regression

In Chapter 1, we introduced a four-step process for organizing a statistics problem. Here is another example of the four-step process in action.

EXAMPLE

Gesell Scores

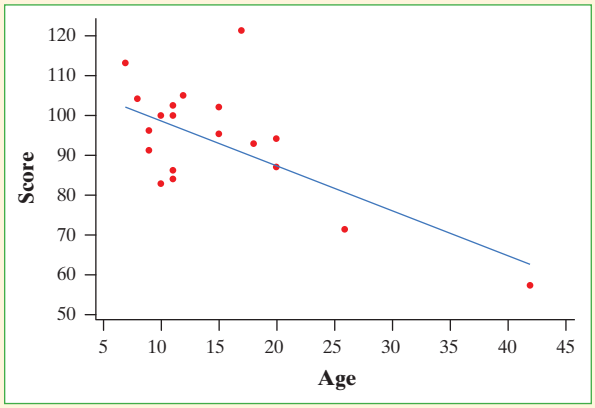
STEP 4

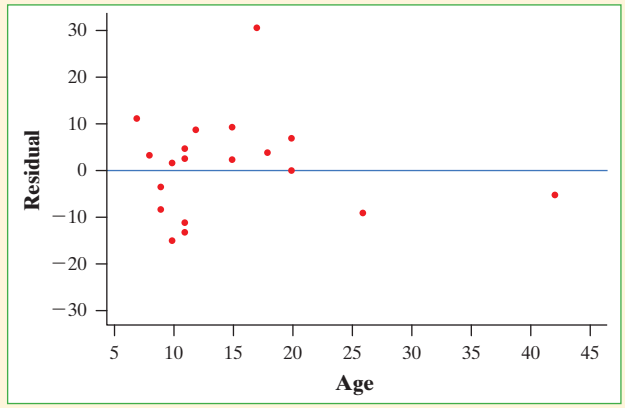
Putting it all together

Does the age at which a child begins to talk predict a later score on a test of mental ability? A study of the development of young children recorded the age in months at which each of 21 children spoke their first word and their Gesell Adaptive Score, the result of an aptitude test taken much later.[16] The data appear in the table below, along with a scatterplot, residual plot, and computer output. Should we use a linear model to predict a child's Gesell score from his or her age at first word? If so, how accurate will our predictions be?

Age (months) at first word and Gesell score

CHILD	AGE	SCORE	CHILD	AGE	SCORE	CHILD	AGE	SCORE
1	15	95	8	11	100	15	11	102
2	26	71	9	8	104	16	10	100
3	10	83	10	20	94	17	12	105
4	9	91	11	7	113	18	42	57
5	15	102	12	9	96	19	17	121
6	20	87	13	10	83	20	11	86
7	18	93	14	11	84	21	10	100





Predictor	Coef	SE Coef	T	P
Constant	109.874	5.068	21.68	0.000
Age	−1.1270	0.3102	−3.63	0.002

S = 11.0229 R-Sq = 41.0% R-Sq(adj) = 37.9%

STATE: Is a linear model appropriate for these data? If so, how well does the least-squares regression line fit the data?

PLAN: To determine whether a linear model is appropriate, we will look at the scatterplot and residual plot to see if the association is linear or nonlinear. Then, if a linear model is appropriate, we will use the standard deviation of the residuals and r^2 to measure how well the least-squares line fits the data.

DO: The scatterplot shows a moderately strong, negative linear association between age at first word and Gesell score. There are a couple of outliers in the scatterplot. Child 19 has a very high Gesell score for his or her age at first word. Also, child 18 didn't speak his or her first word until much later than the other children in the study and has a much lower Gesell score. The residual plot does not have any obvious patterns, confirming what we saw in the scatterplot—a linear model is appropriate for these data.

From the computer output, the equation of the least-squares regression line is $\hat{y} = 109.874 - 1.1270x$. The standard deviation of the residuals is $s = 11.0229$. This means that our predictions will typically be off by 11.0229 points when we use the linear model to predict Gesell scores from age at first word. Finally, 41% of the variation in Gesell score is accounted for by the linear model relating Gesell score to age at first word.

CONCLUDE: Although a linear model is appropriate for these data, our predictions might not be very accurate. Our typical prediction error is about 11 points, and more than half of the variation in Gesell score is still unaccounted for. Furthermore, we should be hesitant to use this model to make predictions until we understand the effect of the two outliers on the regression results.

For Practice *Try Exercise* 67

Correlation and Regression Wisdom

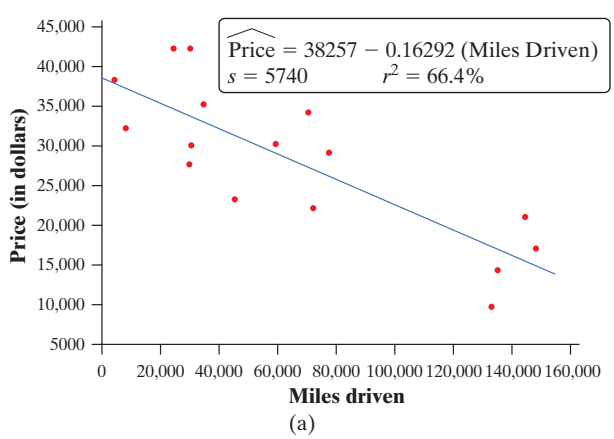
Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, you should be aware of their limitations.

1. *The distinction between explanatory and response variables is important in regression.* Least-squares regression makes the distances of the data points from the line small only in the y direction. If we reverse the roles of the two variables, we get a different least-squares regression line.This isn't true for correlation: switching x and y doesn't affect the value of r.

EXAMPLE Predicting Price, Predicting Miles Driven

Two different regression lines

Figure 3.17(a) repeats the scatterplot of the Ford F-150 data with the least-squares regression line for predicting price from miles driven. We might also use the data on these 16 trucks to predict the number of miles driven from the price of the truck. Now the roles of the variables are reversed: price is the explanatory variable and miles driven is the response variable. Figure 3.17(b) shows a scatterplot of these data with the least-squares regression line for predicting miles driven from price. The two regression lines are very different. The standard deviations of the residuals are different as well. In (a), the standard deviation is $s = 5740$ *dollars*, but in (b) the standard deviation is $s = 28{,}716$ *miles*. However, no matter which variable we put on the x axis, the value of r^2 is 66.4% and the correlation is $r = -0.815$.



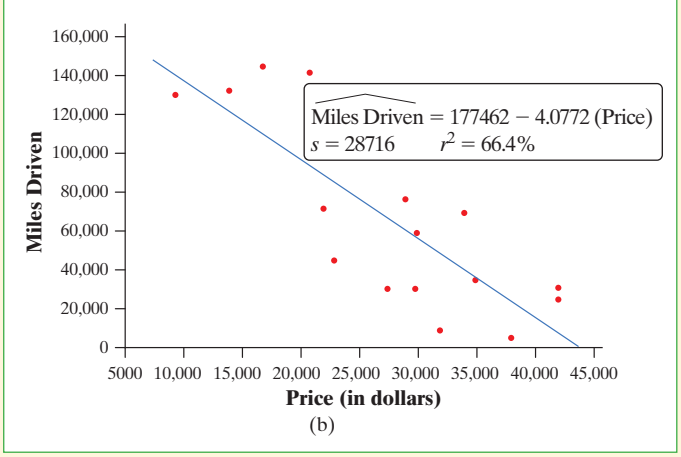
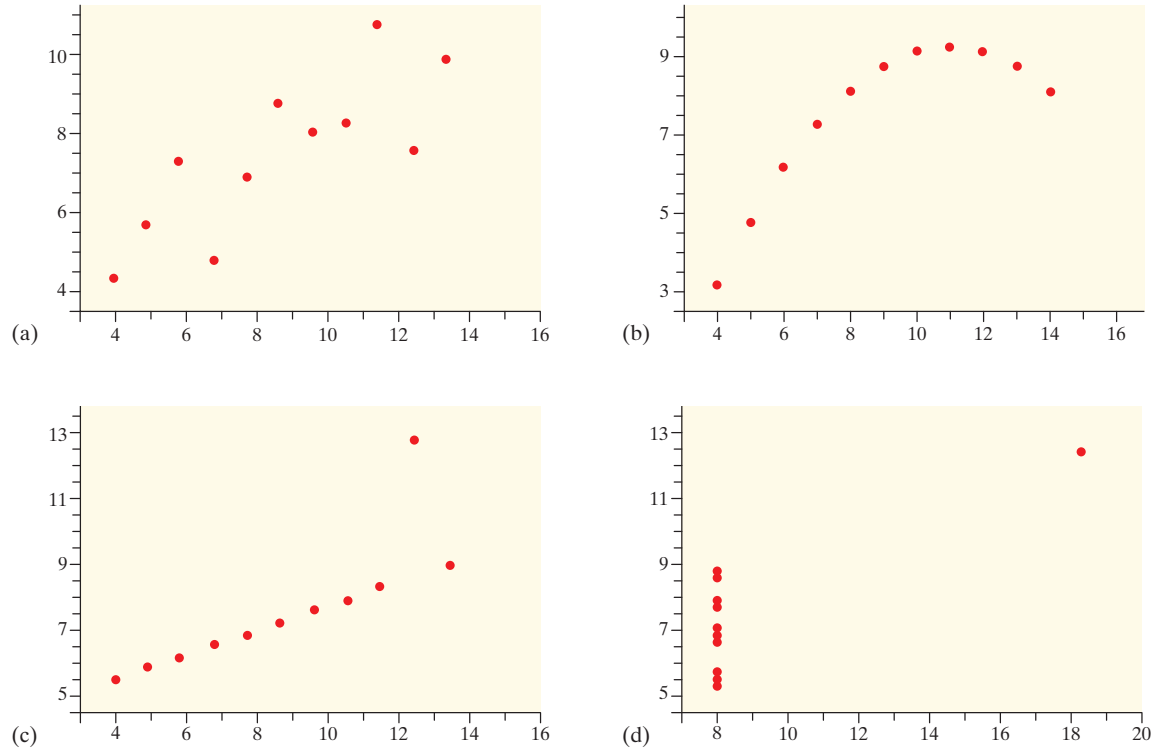


FIGURE 3.17 (a) Scatterplot with least-squares regression line for predicting price from miles driven. (b) Scatterplot with least-squares regression line for predicting miles driven from price.

2. *Correlation and regression lines describe only linear relationships.* You can calculate the correlation and the least-squares line for any relationship between two quantitative variables, but the results are useful only if the scatterplot shows a linear pattern. *Always plot your data!*

The following four scatterplots show very different associations. Which do you think has the highest correlation?



Answer: All four have the same correlation, $r = 0.816$. Furthermore, the least-squares regression line for each association is exactly the same, $\hat{y} = 3 + 0.5x$. These four data sets, developed by statistician Frank Anscombe, illustrate the importance of graphing data before doing calculations.[17]

3. *Correlation and least-squares regression lines are not resistant.* You already know that the correlation r is not resistant. One unusual point in a scatterplot can greatly change the value of r. Is the least-squares line resistant? Not surprisingly, the answer is no.

Let's revisit the age at first word and Gesell score data to shed some light on this issue. The scatterplot and residual plot for these data are shown in Figure 3.18. The two outliers, child 18 and child 19, are indicated on each plot.

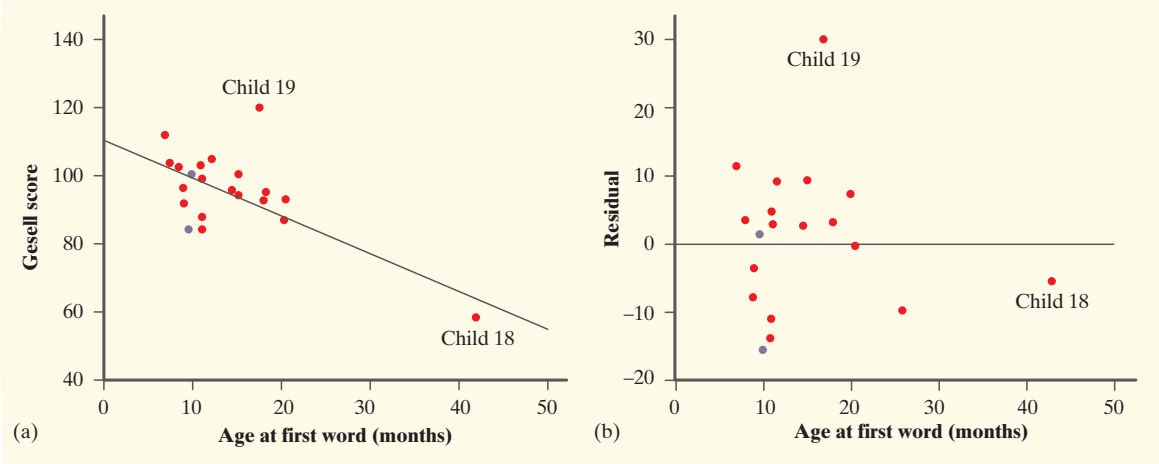


FIGURE 3.18 (a) Scatterplot of Gesell Adaptive Scores versus the age at first word for 21 children. The line is the least-squares regression line for predicting Gesell score from age at first word. (b) Residual plot for the regression. Child 18 and Child 19 are outliers. Each blue point in the graphs stands for two individuals.

Child 19 has a very large residual because this point lies far from the regression line. However, Child 18 has a fairly small residual. That's because Child 18's point is close to the line. How do these two outliers affect the regression?

Figure 3.19 shows the results of removing each of these points on the correlation and the regression line. The graph adds two more regression lines, one calculated after leaving out Child 18 and the other after leaving out Child 19. You can see that removing the point for Child 18 moves the line quite a bit. (In fact, the equation of the new least-squares line is $\hat{y} = 105.630 - 0.779x$.) Because of Child 18's extreme position on the age scale, this point has a strong *influence* on the position of the regression line. However, removing Child 19 has little effect on the regression line.

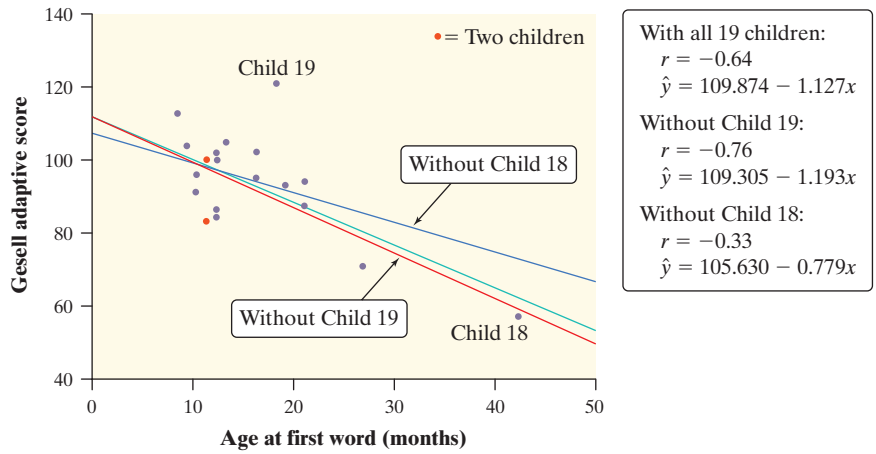


FIGURE 3.19 Three least-squares regression lines of Gesell score on age at first word. The green line is calculated from all the data. The dark blue line is calculated leaving out Child 18. Child 18 is an influential observation because leaving out this point moves the regression line quite a bit. The red line is calculated leaving out only Child 19.

Least-squares lines make the sum of the squares of the vertical distances to the points as small as possible. A point that is extreme in the x direction with no other points near it pulls the line toward itself. We call such points **influential.**

DEFINITION: Outliers and influential observations in regression

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction but not the x direction of a scatterplot have large residuals. Other outliers may not have large residuals.

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the least-squares regression line.

We did not need the distinction between outliers and influential observations in Chapter 1. A single large salary that pulls up the mean salary $\bar{x}$ for a group of workers is an outlier because it lies far above the other salaries. It is also influential, because the mean changes when it is removed. In the regression setting, however, not all outliers are influential. The least-squares line is most likely to be heavily influenced by observations that are outliers in the x direction. The scatterplot will alert you to such observations. Influential points often have small residuals, because they pull the regression line toward themselves. If you look at just a residual plot, you may miss influential points.

The best way to verify that a point is influential is to find the regression line both with and without the unusual point, as in Figure 3.19. If the line moves more than a small amount when the point is deleted, the point is influential.

THINK ABOUT IT

How much difference can one point make? The strong influence of Child 18 makes the original regression of Gesell score on age at first word misleading. The original data have $r^2 = 0.41$. That is, the least-squares line relating age at which a child begins to talk with Gesell score explains 41% of the variation on this later test of mental ability. This relationship is strong enough to be interesting to parents. If we leave out Child 18, r^2 drops to only 11%. The apparent strength of the association was largely due to a single influential observation.

What should the child development researcher do? She must decide whether Child 18 is so slow to speak that this individual should not be allowed to influence the analysis. If she excludes Child 18, much of the evidence for a connection between the age at which a child begins to talk and later ability score vanishes. If she keeps Child 18, she needs data on other children who were also slow to begin talking, so that the analysis no longer depends so heavily on just one child.

We finish with our most important caution about correlation and regression.

4. *Association does not imply causation.* When we study the relationship between two variables, we often hope to show that changes in the explanatory variable *cause* changes in the response variable. *A strong association between two variables is not enough to draw conclusions about cause and effect.* Sometimes an observed association really does reflect cause and effect. A household that heats with natural gas uses more gas in colder months because cold weather requires burning more gas to stay warm. In other cases, an association is explained by other variables, and the conclusion that x causes y is not valid.

EXAMPLE

Does Having More Cars Make You Live Longer?

Association, not causation

A serious study once found that people with two cars live longer than people who own only one car.[18] Owning three cars is even better, and so on. There is a substantial positive association between number of cars x and length of life y.

The basic meaning of causation is that by changing x, we can bring about a change in y. Could we lengthen our lives by buying more cars? No. The study used number of cars as a quick indicator of wealth. Well-off people tend to have more cars. They also tend to live longer, probably because they are better educated, take better care of themselves, and get better medical care. The cars have nothing to do with it. There is no cause-and-effect link between number of cars and length of life.

Associations such as those in the previous example are sometimes called "nonsense associations." The association is real. What is nonsense is the conclusion that changing one of the variables causes changes in the other. Another variable—such as personal wealth in this example—that influences both x and y can create a strong association even though there is no direct connection between x and y.

Remember: It only makes sense to talk about the *correlation* between two *quantitative* variables. If one or both variables are categorical, you should refer to the *association* between the two variables. To be safe, you can use the more general term "association" when describing the relationship between any two variables.

ASSOCIATION DOES NOT IMPLY CAUSATION

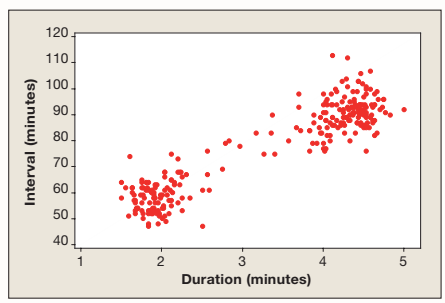
An association between an explanatory variable x and a response variable y, even if it is very strong, is not by itself good evidence that changes in x actually cause changes in y.

Here is a chance to use the skills you have gained to address the question posed at the beginning of the chapter.

case closed

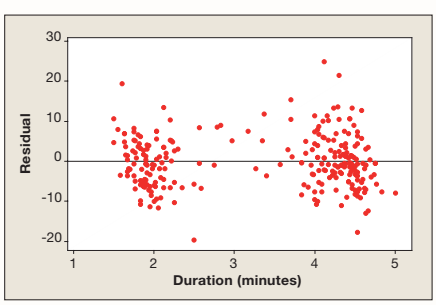
How Faithful Is Old Faithful?

In the chapter-opening Case Study (page 141), the Starnes family had just missed seeing Old Faithful erupt. They wondered how long it would be until the next eruption. The scatterplot below shows data on the duration (in minutes) and the interval of time until the next eruption (also in minutes) for each Old Faithful eruption in the month before their visit.



1. Describe the nature of the relationship between interval and duration.

Here is some computer output from a least-squares regression analysis on these data.



Regression Analysis: Interval versus Duration

Predictor	Coef	SE Coef	T	P
Constant	33.347	1.201	27.76	0.000
Duration	13.2854	0.3404	39.03	0.000

S = 6.49336 R-Sq = 85.4% R-Sq(adj) = 85.3%

2. Is a linear model appropriate? Justify your answer.
3. Give the equation of the least-squares regression line. Be sure to define any variables you use.
4. Park rangers indicated that the eruption of Old Faithful that just finished lasted 3.9 minutes. How long do you predict the Starnes family will have to wait for the next eruption? Show how you arrived at your answer.
5. The actual time that the Starnes family has to wait is probably not exactly equal to your prediction in Question 4. Based on the computer output, about how far off do you expect the prediction to be? Explain.

Section 3.2 Summary

- A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. You can use a regression line to **predict** the value of y for any value of x by substituting this x into the equation of the line.
- The **slope** b of a regression line $\hat{y} = a + bx$ is the rate at which the predicted response $\hat{y}$ changes along the line as the explanatory variable x changes. Specifically, b is the *predicted* change in y when x increases by 1 unit.
- The **y intercept** a of a regression line $\hat{y} = a + bx$ is the predicted response $\hat{y}$ when the explanatory variable x equals 0. This prediction is of no statistical use unless x can actually take values near 0.
- Avoid **extrapolation,** the use of a regression line for prediction using values of the explanatory variable outside the range of the data from which the line was calculated.
- The most common method of fitting a line to a scatterplot is least squares. The **least-squares regression line** is the straight line $\hat{y} = a + bx$ that minimizes the sum of the squares of the vertical distances of the observed points from the line.
- You can examine the fit of a regression line by studying the **residuals,** which are the differences between the observed and predicted values of y. Be on the lookout for patterns in the **residual plot**, which indicate that a linear model may not be appropriate.
- The **standard deviation of the residuals** s measures the typical size of the prediction errors (residuals) when using the regression line.
- The **coefficient of determination** r^2 is the fraction of the variation in the response variable that is accounted for by least-squares regression on the explanatory variable.
- The least-squares regression line of y on x is the line with slope $b = r(s_y/s_x)$ and intercept $a = \bar{y} - b\bar{x}$. This line always passes through the point $(\bar{x}, \bar{y})$.
- Correlation and regression must be interpreted with caution. Plot the data to be sure that the relationship is roughly linear and to detect **outliers.** Also look for **influential observations,** individual points that substantially change the correlation or the regression line. Outliers in x are often influential for the regression line.
- Most of all, be careful not to conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated.

3.2 TECHNOLOGY CORNERS

TI-Nspire Instructions in Appendix B; HP Prime instructions on the book's Web site.

Section 3.2 Exercises

35. **What's my line?** You use the same bar of soap to shower each morning. The bar weighs 80 grams when it is new. Its weight goes down by 6 grams per day on average. What is the equation of the regression line for predicting weight from days of use?

36. **What's my line?** An eccentric professor believes that a child with IQ 100 should have a reading test score of 50 and predicts that reading score should increase by 1 point for every additional point of IQ. What is the equation of the professor's regression line for predicting reading score from IQ?

37. **Gas mileage** We expect a car's highway gas mileage to be related to its city gas mileage. Data for all 1198 vehicles in the government's recent *Fuel Economy Guide* give the regression line: predicted highway mpg = 4.62 + 1.109 (city mpg).

(a) What's the slope of this line? Interpret this value in context.

(b) What's the y intercept? Explain why the value of the intercept is not statistically meaningful.

(c) Find the predicted highway mileage for a car that gets 16 miles per gallon in the city.

38. **IQ and reading scores** Data on the IQ test scores and reading test scores for a group of fifth-grade children give the following regression line: predicted reading score = −33.4 + 0.882(IQ score).

(a) What's the slope of this line? Interpret this value in context.

(b) What's the y intercept? Explain why the value of the intercept is not statistically meaningful.

(c) Find the predicted reading score for a child with an IQ score of 90.

39. **Acid rain** Researchers studying acid rain measured the acidity of precipitation in a Colorado wilderness area for 150 consecutive weeks. Acidity is measured by pH. Lower pH values show higher acidity. The researchers observed a linear pattern over time. They reported that the regression line $\widehat{pH} = 5.43 - 0.0053(\text{weeks})$ fit the data well.[19]

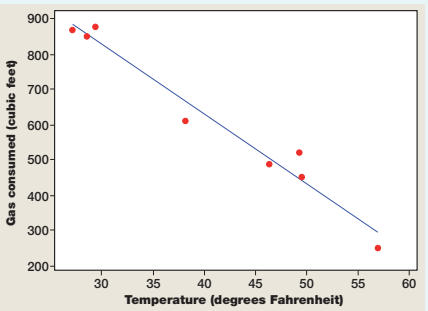
pg 166

(a) Identify the slope of the line and explain what it means in this setting.

(b) Identify the y intercept of the line and explain what it means in this setting.

(c) According to the regression line, what was the pH at the end of this study?

40. **How much gas?** In Exercise 4 (page 159), we examined the relationship between the average monthly temperature and the amount of natural gas consumed in Joan's midwestern home. The figure below shows the original scatterplot with the least-squares line added. The equation of the least-squares line is $\hat{y} = 1425 - 19.87x$.



(a) Identify the slope of the line and explain what it means in this setting.

(b) Identify the y intercept of the line. Explain why it's risky to use this value as a prediction.

(c) Use the regression line to predict the amount of natural gas Joan will use in a month with an average temperature of 30°F.

41. **Acid rain** Refer to Exercise 39. Would it be appropriate to use the regression line to predict pH after 1000 months? Justify your answer.

42. **How much gas?** Refer to Exercise 40. Would it be appropriate to use the regression line to predict Joan's natural-gas consumption in a future month with an average temperature of 65°F? Justify your answer.

43. **Least-squares idea** The table below gives a small set of data. Which of the following two lines fits the data better: $\hat{y} = 1 - x$ or $\hat{y} = 3 - 2x$? Use the least-squares criterion to justify your answer. (*Note:* Neither of these two lines is the least-squares regression line for these data.)

x:	−1	1	1	3	5
y:	2	0	1	−1	−5

44. **Least-squares idea** In Exercise 40, the line drawn on the scatterplot is the least-squares regression line. Explain the meaning of the phrase "least-squares" to Joan, who knows very little about statistics.

45. **Acid rain** In the acid rain study of Exercise 39, the actual pH measurement for Week 50 was 5.08. Find and interpret the residual for this week.

pg 169

46. **How much gas?** Refer to Exercise 40. During March, the average temperature was 46.4°F and Joan used 490 cubic feet of gas per day. Find and interpret the residual for this month.

47. **Bird colonies** Exercise 6 (page 159) examined the relationship between the number of new birds y and percent of returning birds x for 13 sparrowhawk colonies. Here are the data once again.

Percent return:	74	66	81	52	73	62	52	45	62	46	60	46	38
New adults:	5	6	8	11	12	15	16	17	18	18	19	20	20

(a) Use your calculator to help make a scatterplot.

(b) Use your calculator's regression function to find the equation of the least-squares regression line. Add this line to your scatterplot from (a).

(c) Explain in words what the slope of the regression line tells us.

(d) Calculate and interpret the residual for the colony that had 52% of the sparrowhawks return and 11 new adults.

48. **Do heavier people burn more energy?** Exercise 10 (page 160) presented data on the lean body mass and resting metabolic rate for 12 women who were subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate, in calories burned per 24 hours, is the rate at which the body consumes energy. Here are the data again.

Mass:	36.1	54.6	48.5	42.0	50.6	42.0	40.3	33.1	42.4	34.5	51.1	41.2
Rate:	995	1425	1396	1418	1502	1256	1189	913	1124	1052	1347	1204

(a) Use your calculator to help make a scatterplot.

(b) Use your calculator's regression function to find the equation of the least-squares regression line. Add this line to your scatterplot from part (a).

(c) Explain in words what the slope of the regression line tells us.

(d) Calculate and interpret the residual for the woman who had a lean body mass of 50.6 kg and a metabolic rate of 1502.

49. **Bird colonies** Refer to Exercise 47.

(a) Use your calculator to make a residual plot. Describe what this graph tells you about the appropriateness of using a linear model.

(b) Which point has the largest residual? Explain what this residual means in context.

50. **Do heavier people burn more energy?** Refer to Exercise 48.

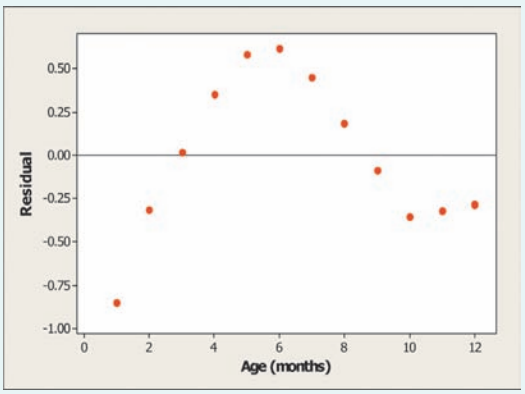
(a) Use your calculator to make a residual plot. Describe what this graph tells you about the appropriateness of using a linear model.

(b) Which point has the largest residual? Explain what the value of that residual means in context.

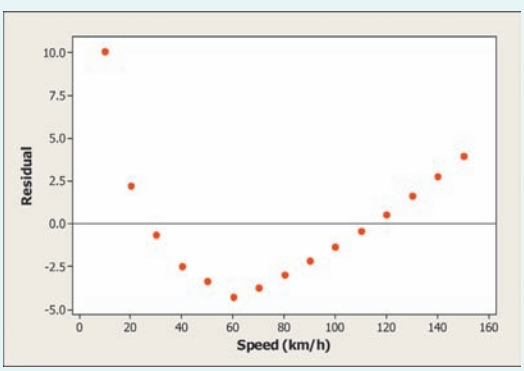
51. **Nahya infant weights** A study of nutrition in developing countries collected data from the Egyptian village of Nahya. Here are the mean weights (in kilograms) for 170 infants in Nahya who were weighed each month during their first year of life:

Age (months):	1	2	3	4	5	6	7	8	9	10	11	12
Weight (kg):	4.3	5.1	5.7	6.3	6.8	7.1	7.2	7.2	7.2	7.2	7.5	7.8

A hasty user of statistics enters the data into software and computes the least-squares line without plotting the data. The result is $\widehat{\text{weight}} = 4.88 + 0.267\,(\text{age})$. A residual plot is shown below. Would it be appropriate to use this regression line to predict y from x? Justify your answer.

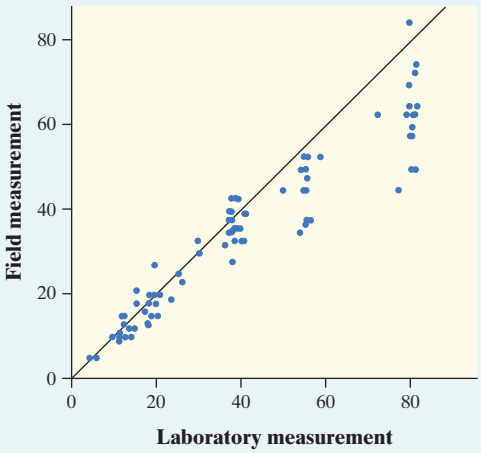


52. **Driving speed and fuel consumption** Exercise 9 (page 160) gives data on the fuel consumption y of a car at various speeds x. Fuel consumption is measured in liters of gasoline per 100 kilometers driven and speed is measured in kilometers per hour. A statistical software package gives the least-squares regression line and the residual plot shown below. The regression line is $\hat{y} = 11.058 - 0.01466x$. Would it be appropriate to use the regression line to predict y from x? Justify your answer.



53. **Oil and residuals** The Trans-Alaska Oil Pipeline is a tube that is formed from 1/2-inch-thick steel and that carries oil across 800 miles of sensitive arctic and subarctic terrain. The pipe segments and the welds that join them were carefully examined before installation. How accurate are field measurements of the depth of small defects? The figure below compares the results of measurements on 100 defects made in the field with

measurements of the same defects made in the laboratory.[20] The line $y = x$ is drawn on the scatterplot.

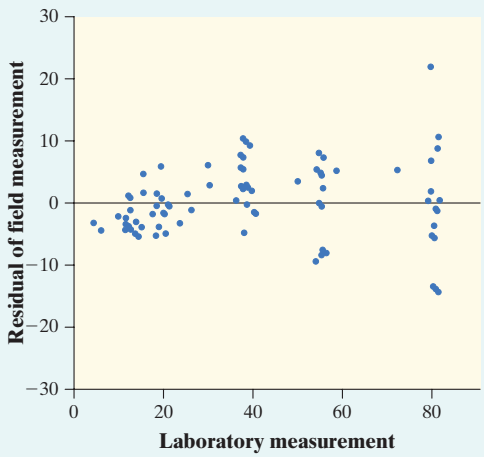


(a) Describe the overall pattern you see in the scatterplot, as well as any deviations from that pattern.

(b) If field and laboratory measurements all agree, then the points should fall on the $y = x$ line drawn on the plot, except for small variations in the measurements. Is this the case? Explain.

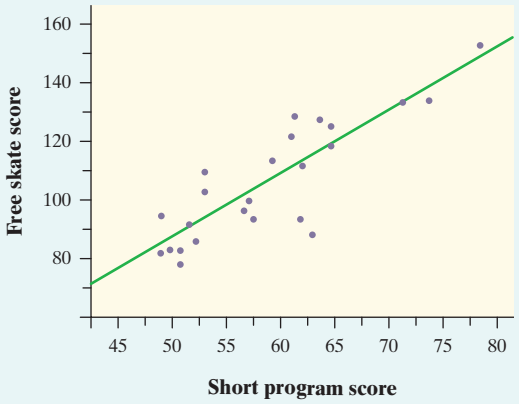
(c) The line drawn on the scatterplot ($y = x$) is *not* the least-squares regression line. How would the slope and y intercept of the least-squares line compare? Justify your answer.

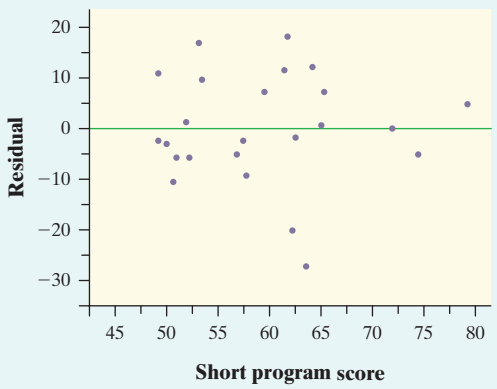
54. Oil and residuals Refer to Exercise 53. The following figure shows a residual plot for the least-squares regression line. Discuss what the residual plot tells you about the appropriateness of using a linear model.



55. Olympic figure skating For many people, the women's figure skating competition is the highlight of the Olympic Winter Games. Scores in the short program x and scores in the free skate y were recorded for each of the 24 skaters who competed in both rounds during the 2010 Winter Olympics in Vancouver, Canada.[21] A regression analysis was performed using these data. The scatterplot and residual plot follow. The equation of the least-squares regression line is $\hat{y} = -16.2 + 2.07x$. Also, $s = 10.2$ and $r^2 = 0.736$.

pg 180





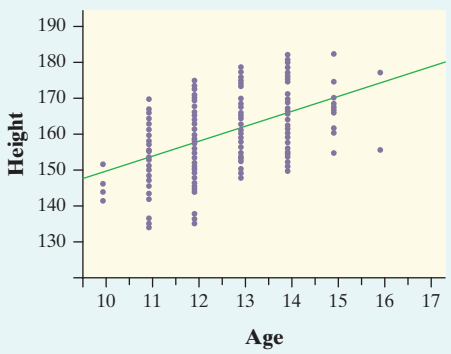
(a) Calculate and interpret the residual for the gold medal winner, Yu-Na Kim, who scored 78.50 in the short program and 150.06 in the free skate.

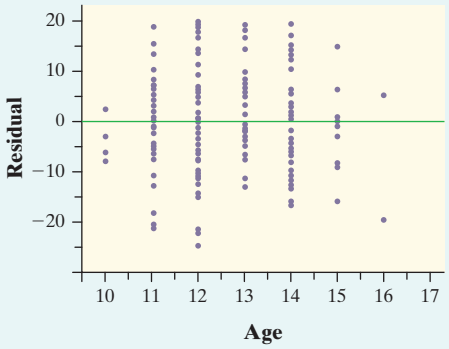
(b) Is a linear model appropriate for these data? Explain.

(c) Interpret the value of s.

(d) Interpret the value of r^2.

56. Age and height A random sample of 195 students was selected from the United Kingdom using the CensusAtSchool data selector. The age (in years) x and height (in centimeters) y was recorded for each of the students. A regression analysis was performed using these data. The scatterplot and residual plot are shown below. The equation of the least-squares regression line is $\hat{y} = 106.1 + 4.21x$. Also, $s = 8.61$ and $r^2 = 0.274$.





(a) Calculate and interpret the residual for the student who was 141 cm tall at age 10.

(b) Is a linear model appropriate for these data? Explain.

(c) Interpret the value of s.

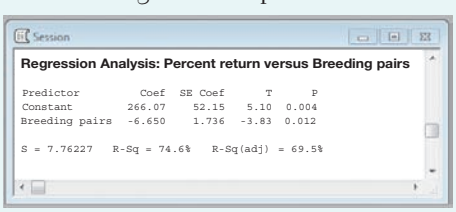
(d) Interpret the value of r^2.

57. **Bird colonies** Refer to Exercises 47 and 49. For the regression you performed earlier, $r^2 = 0.56$ and $s = 3.67$. Explain what each of these values means in this setting.

58. **Do heavier people burn more energy?** Refer to Exercises 48 and 50. For the regression you performed earlier, $r^2 = 0.768$ and $s = 95.08$. Explain what each of these values means in this setting.

59. **Merlins breeding** Exercise 13 (page 160) gives data on the number of breeding pairs of merlins in an isolated area in each of seven years and the percent of males who returned the next year. The data show that the percent returning is lower after successful breeding seasons and that the relationship is roughly linear. The figure below shows Minitab regression output for these data.

pg 181



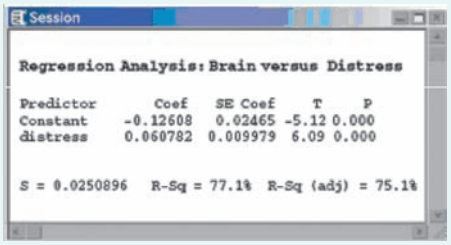
(a) What is the equation of the least-squares regression line for predicting the percent of males that return from the number of breeding pairs? Use the equation to predict the percent of returning males after a season with 30 breeding pairs.

(b) What percent of the year-to-year variation in percent of returning males is accounted for by the straight-line relationship with number of breeding pairs the previous year?

(c) Use the information in the figure to find the correlation r between percent of males that return and number of breeding pairs. How do you know whether the sign of r is + or −?

(d) Interpret the value of s in this setting.

60. **Does social rejection hurt?** Exercise 14 (page 161) gives data from a study that shows that social exclusion causes "real pain." That is, activity in an area of the brain that responds to physical pain goes up as distress from social exclusion goes up. A scatterplot shows a moderately strong, linear relationship. The figure below shows Minitab regression output for these data.



(a) What is the equation of the least-squares regression line for predicting brain activity from social distress score? Use the equation to predict brain activity for social distress score 2.0.

(b) What percent of the variation in brain activity among these subjects is accounted for by the straight-line relationship with social distress score?

(c) Use the information in the figure to find the correlation r between social distress score and brain activity. How do you know whether the sign of r is + or −?

(d) Interpret the value of s in this setting.

61. **Husbands and wives** The mean height of married American women in their early twenties is 64.5 inches and the standard deviation is 2.5 inches. The mean height of married men the same age is 68.5 inches, with standard deviation 2.7 inches. The correlation between the heights of husbands and wives is about $r = 0.5$.

pg 183

(a) Find the equation of the least-squares regression line for predicting a husband's height from his wife's height for married couples in their early 20s. Show your work.

(b) Suppose that the height of a randomly selected wife was 1 standard deviation below average. Predict the height of her husband *without* using the least-squares line. Show your work.

62. **The stock market** Some people think that the behavior of the stock market in January predicts its behavior for the rest of the year. Take the explanatory variable x to be the percent change in a stock market index in January and the response variable y to be the change in the index for the entire year. We expect a positive correlation between x and y because the change during January contributes to the full year's change. Calculation from data for an 18-year period gives

$$\bar{x} = 1.75\% \quad s_x = 5.36\% \quad \bar{y} = 9.07\%$$
$$s_y = 15.35\% \quad r = 0.596$$

(a) Find the equation of the least-squares line for predicting full-year change from January change. Show your work.

(b) Suppose that the percent change in a particular January was 2 standard deviations above average. Predict the percent change for the entire year, *without* using the least-squares line. Show your work.

63. Husbands and wives Refer to Exercise 61.

(a) Find r^2 and interpret this value in context.

(b) For these data, $s = 1.2$. Interpret this value.

64. The stock market Refer to Exercise 62.

(a) Find r^2 and interpret this value in context.

(b) For these data, $s = 8.3$. Interpret this value.

65. Will I bomb the final? We expect that students who do well on the midterm exam in a course will usually also do well on the final exam. Gary Smith of Pomona College looked at the exam scores of all 346 students who took his statistics class over a 10-year period.[22] Assume that both the midterm and final exam were scored out of 100 points.

(a) State the equation of the least-squares regression line if each student scored the same on the midterm and the final.

(b) The actual least-squares line for predicting final-exam score y from midterm-exam score x was $\hat{y} = 46.6 + 0.41x$. Predict the score of a student who scored 50 on the midterm and a student who scored 100 on the midterm.

(c) Explain how your answers to part (b) illustrate regression to the mean.

66. It's still early We expect that a baseball player who has a high batting average in the first month of the season will also have a high batting average the rest of the season. Using 66 Major League Baseball players from the 2010 season,[23] a least-squares regression line was calculated to predict rest-of-season batting average y from first-month batting average x. *Note:* A player's batting average is the proportion of times at bat that he gets a hit. A batting average over 0.300 is considered very good in Major League Baseball.

(a) State the equation of the least-squares regression line if each player had the same batting average the rest of the season as he did in the first month of the season.

(b) The actual equation of the least-squares regression line is $\hat{y} = 0.245 + 0.109x$. Predict the rest-of-season batting average for a player who had a 0.200 batting average the first month of the season and for a player who had a 0.400 batting average the first month of the season.

(c) Explain how your answers to part (b) illustrate regression to the mean.

67. Beavers and beetles Do beavers benefit beetles? Researchers laid out 23 circular plots, each 4 meters in diameter, in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Ecologists think that the new sprouts from stumps are more tender than other cottonwood growth, so that beetles prefer them. If so, more stumps should produce more beetle larvae. Here are the data:[24]

Stumps:	2	2	1	3	3	4	3	1	2	5	1	3
Beetle larvae:	10	30	12	24	36	40	43	11	27	56	18	40
Stumps:	2	1	2	2	1	1	4	1	2	1	4	
Beetle larvae:	25	8	21	14	16	6	54	9	13	14	50	

Can we use a linear model to predict the number of beetle larvae from the number of stumps? If so, how accurate will our predictions be? Follow the four-step process.

68. Fat and calories The number of calories in a food item depends on many factors, including the amount of fat in the item. The data below show the amount of fat (in grams) and the number of calories in 7 beef sandwiches at McDonalds.[25]

Sandwich	Fat	Calories
Big Mac®	29	550
Quarter Pounder® with Cheese	26	520
Double Quarter Pounder® with Cheese	42	750
Hamburger	9	250
Cheeseburger	12	300
Double Cheeseburger	23	440
McDouble	19	390

Can we use a linear model to predict the number of calories from the amount of fat? If so, how accurate will our predictions be? Follow the four-step process.

69. Managing diabetes People with diabetes measure their fasting plasma glucose (FPG; measured in units of milligrams per milliliter) after fasting for at least 8 hours. Another measurement, made at regular medical checkups, is called HbA. This is roughly the percent of red blood cells that have a glucose molecule attached. It measures average exposure to glucose over a period of several months. The table below gives data on both HbA and FPG for 18 diabetics five months after they had completed a diabetes education class.[27]

Subject	HbA (%)	FPG (mg/mL)	Subject	HbA (%)	FPG (mg/mL)
1	6.1	141	10	8.7	172
2	6.3	158	11	9.4	200
3	6.4	112	12	10.4	271
4	6.8	153	13	10.6	103
5	7.0	134	14	10.7	172
6	7.1	95	15	10.7	359
7	7.5	96	16	11.2	145
8	7.7	78	17	13.7	147
9	7.9	148	18	19.3	255

(a) Make a scatterplot with HbA as the explanatory variable. Describe what you see.

(b) Subject 18 is an outlier in the x direction. What effect do you think this subject has on the correlation? What effect do you think this subject has on the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this subject to confirm your answer.

(c) Subject 15 is an outlier in the y direction. What effect do you think this subject has on the correlation? What effect do you think this subject has on the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this subject to confirm your answer.

70. **Rushing for points** What is the relationship between rushing yards and points scored in the 2011 National Football League? The table below gives the number of rushing yards and the number of points scored for each of the 16 games played by the 2011 Jacksonville Jaguars.[26]

Game	Rushing yards	Points scored
1	163	16
2	112	3
3	128	10
4	104	10
5	96	20
6	133	13
7	132	12
8	84	14
9	141	17
10	108	10
11	105	13
12	129	14
13	116	41
14	116	14
15	113	17
16	190	19

(a) Make a scatterplot with rushing yards as the explanatory variable. Describe what you see.

(b) The number of rushing yards in Game 16 is an outlier in the x direction. What effect do you think this game has on the correlation? On the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this game to confirm your answers.

(c) The number of points scored in Game 13 is an outlier in the y direction. What effect do you think this game has on the correlation? On the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this game to confirm your answers.

Multiple choice: Select the best answer for Exercises 71 to 78.

71. Which of the following is *not* a characteristic of the least-squares regression line?

(a) The slope of the least-squares regression line is always between -1 and 1.

(b) The least-squares regression line always goes through the point $(\bar{x}, \bar{y})$.

(c) The least-squares regression line minimizes the sum of squared residuals.

(d) The slope of the least-squares regression line will always have the same sign as the correlation.

(e) The least-squares regression line is not resistant to outliers.

72. Each year, students in an elementary school take a standardized math test at the end of the school year. For a class of fourth-graders, the average score was 55.1 with a standard deviation of 12.3. In the third grade, these same students had an average score of 61.7 with a standard deviation of 14.0. The correlation between the two sets of scores is $r = 0.95$. Calculate the equation of the least-squares regression line for predicting a fourth-grade score from a third-grade score.

(a) $\hat{y} = 3.60 + 0.835x$

(b) $\hat{y} = 15.69 + 0.835x$

(c) $\hat{y} = 2.19 + 1.08x$

(d) $\hat{y} = -11.54 + 1.08x$

(e) Cannot be calculated without the data.

73. Using data from the 2009 LPGA tour, a regression analysis was performed using x = average driving distance and y = scoring average. Using the output from the regression analysis shown below, determine the equation of the least-squares regression line.

```
Predictor          Coef        SE Coef    T       P
Constant           87.974      2.391      36.78   0.000
Driving Distance   −0.060934   0.009536   −6.39   0.000

S = 1.01216    R-Sq = 22.1%    R-Sq(adj) = 21.6%
```

(a) $\hat{y} = 87.947 + 2.391x$

(b) $\hat{y} = 87.947 + 1.01216x$

(c) $\hat{y} = 87.947 - 0.060934x$

(d) $\hat{y} = -0.060934 + 1.01216x$

(e) $\hat{y} = -0.060934 + 87.947x$

Exercises 74 to 78 refer to the following setting. Measurements on young children in Mumbai, India, found this least-squares line for predicting height y from arm span x:[28]

$$\hat{y} = 6.4 + 0.93x$$

Measurements are in centimeters (cm).

74. By looking at the equation of the least-squares regression line, you can see that the correlation between height and arm span is

(a) greater than zero.

(b) less than zero.

(c) 0.93.

(d) 6.4.

(e) Can't tell without seeing the data.

75. In addition to the regression line, the report on the Mumbai measurements says that $r^2 = 0.95$. This suggests that

(a) although arm span and height are correlated, arm span does not predict height very accurately.

(b) height increases by $\sqrt{0.95} = 0.97$ cm for each additional centimeter of arm span.

(c) 95% of the relationship between height and arm span is accounted for by the regression line.

(d) 95% of the variation in height is accounted for by the regression line.

(e) 95% of the height measurements are accounted for by the regression line.

76. One child in the Mumbai study had height 59 cm and arm span 60 cm. This child's residual is

(a) −3.2 cm. (c) −1.3 cm. (e) 62.2 cm.

(b) −2.2 cm. (d) 3.2 cm.

77. Suppose that a tall child with arm span 120 cm and height 118 cm was added to the sample used in this study. What effect will adding this child have on the correlation and the slope of the least-squares regression line?

(a) Correlation will increase, slope will increase.

(b) Correlation will increase, slope will stay the same.

(c) Correlation will increase, slope will decrease.

(d) Correlation will stay the same, slope will stay the same.

(e) Correlation will stay the same, slope will increase.

78. Suppose that the measurements of arm span and height were converted from centimeters to meters by dividing each measurement by 100. How will this conversion affect the values of r^2 and s?

(a) r^2 will increase, s will increase.

(b) r^2 will increase, s will stay the same.

(c) r^2 will increase, s will decrease.

(d) r^2 will stay the same, s will stay the same.

(e) r^2 will stay the same, s will decrease.

Exercises 79 and 80 refer to the following setting.

In its recent *Fuel Economy Guide*, the Environmental Protection Agency gives data on 1152 vehicles. There are a number of outliers, mainly vehicles with very poor gas mileage. If we ignore the outliers, however, the combined city and highway gas mileage of the other 1120 or so vehicles is approximately Normal with mean 18.7 miles per gallon (mpg) and standard deviation 4.3 mpg.

79. **In my Chevrolet** (2.2) The Chevrolet Malibu with a four-cylinder engine has a combined gas mileage of 25 mpg. What percent of all vehicles have worse gas mileage than the Malibu?

80. **The top 10%** (2.2) How high must a vehicle's gas mileage be in order to fall in the top 10% of all vehicles? (The distribution omits a few high outliers, mainly hybrid gas-electric vehicles.)

81. **Marijuana and traffic accidents** (1.1) Researchers in New Zealand interviewed 907 drivers at age 21. They had data on traffic accidents and they asked the drivers about marijuana use. Here are data on the numbers of accidents caused by these drivers at age 19, broken down by marijuana use at the same age:[29]

	Marijuana use per year			
	Never	1–10 times	11–50 times	51 + times
Drivers	452	229	70	156
Accidents caused	59	36	15	50

(a) Make a graph that displays the accident rate for each class. Is there evidence of an association between marijuana use and traffic accidents?

(b) Explain why we can't conclude that marijuana use *causes* accidents.

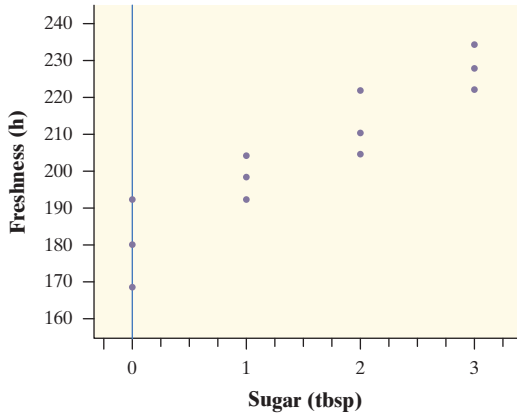
FRAPPY! Free Response AP® Problem, Yay!

The following problem is modeled after actual AP® Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

Two statistics students went to a flower shop and randomly selected 12 carnations. When they got home, the students prepared 12 identical vases with exactly the same amount of water in each vase. They put one tablespoon of sugar in 3 vases, two tablespoons of sugar in 3 vases, and three tablespoons of sugar in 3 vases. In the remaining 3 vases, they put no sugar. After the vases were prepared, the students randomly assigned 1 carnation to each vase

and observed how many hours each flower continued to look fresh. A scatterplot of the data is shown below.



(a) Briefly describe the association shown in the scatterplot.

(b) The equation of the least-squares regression line for these data is $\hat{y} = 180.8 + 15.8x$. Interpret the slope of the line in the context of the study.

(c) Calculate and interpret the residual for the flower that had 2 tablespoons of sugar and looked fresh for 204 hours.

(d) Suppose that another group of students conducted a similar experiment using 12 flowers, but included different varieties in addition to carnations. Would you expect the value of r^2 for the second group's data to be greater than, less than, or about the same as the value of r^2 for the first group's data? Explain.

After you finish, you can view two example solutions on the book's Web site (www.whfreeman.com/tps5e). Determine whether you think each solution is "complete," "substantial," "developing," or "minimal." If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

Chapter Review

Section 3.1: Scatterplots and Correlation

In this section, you learned how to explore the relationship between two quantitative variables. As with distributions of a single variable, the first step is always to make a graph. A scatterplot is the appropriate type of graph to investigate associations between two quantitative variables. To describe a scatterplot, be sure to discuss four characteristics: direction, form, strength, and outliers. The direction of an association might be positive, negative, or neither. The form of an association can be linear or nonlinear. An association is strong if it closely follows a specific form. Finally, outliers are any points that clearly fall outside the pattern of the rest of the data.

The correlation r is a numerical summary that describes the direction and strength of a linear association. When $r > 0$, the association is positive, and when $r < 0$, the association is negative. The correlation will always take values between -1 and 1, with $r = -1$ and $r = 1$ indicating a perfectly linear relationship. Strong linear associations have correlations near 1 or -1, while weak linear relationships have correlations near 0. However, it isn't possible to determine the form of an association from only the correlation. Strong nonlinear relationships can have a correlation close to 1 or a correlation close to 0, depending on the association. You also learned that outliers can greatly affect the value of the correlation and that correlation does not imply causation. That is, we can't assume that changes in one variable cause changes in the other variable, just because they have a correlation close to 1 or –1.

Section 3.2: Least-Squares Regression

In this section, you learned how to use least-squares regression lines as models for relationships between variables that have a linear association. It is important to understand the difference between the actual data and the model used to describe the data. For example, when you are interpreting the slope of a least-squares regression line, describe the *predicted* change in the y variable. To emphasize that the model only provides predicted values, least-squares regression lines are always expressed in terms of $\hat{y}$ instead of y.

The difference between the observed value of y and the predicted value of y is called a residual. Residuals are the key to understanding almost everything in this section. To find the equation of the least-squares regression line, find the line that minimizes the sum of the squared residuals. To see if a linear model is appropriate, make a residual plot. If there is no leftover pattern in the residual plot, you know the model is appropriate. To assess how well a line fits the data, calculate the standard deviation of the residuals s to estimate the size of a typical prediction error. You can also calculate r^2, which measures the fraction of the variation in the y variable that is accounted for by its linear relationship with the x variable.

You also learned how to obtain the equation of a least-squares regression line from computer output and from summary statistics (the means and standard deviations of two variables and their correlation). As with the correlation, the equation of the least-squares regression line and the values of s and r^2 can be greatly influenced by outliers, so be sure to plot the data and note any unusual values before making any calculations.

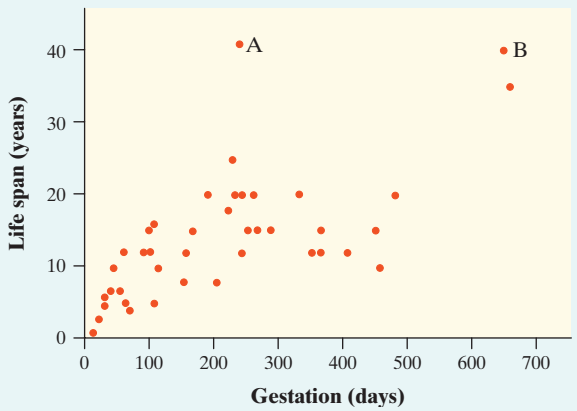
What Did You Learn?

Learning Objective	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
Identify explanatory and response variables in situations where one variable helps to explain or influences the other.	3.1	144	R3.4
Make a scatterplot to display the relationship between two quantitative variables.	3.1	145, 148	R3.4
Describe the direction, form, and strength of a relationship displayed in a scatterplot and recognize outliers in a scatterplot.	3.1	147, 148	R3.1
Interpret the correlation.	3.1	152	R3.3, R3.4
Understand the basic properties of correlation, including how the correlation is influenced by outliers.	3.1	152, 156, 157	R3.1, R3.2
Use technology to calculate correlation.	3.1	Activity on 152, 171	R3.4
Explain why association does not imply causation.	3.1	Discussion on 156, 190	R3.6
Interpret the slope and y intercept of a least-squares regression line.	3.2	166	R3.2, R3.4
Use the least-squares regression line to predict y for a given x. Explain the dangers of extrapolation.	3.2	167, Discussion on 168 (for extrapolation)	R3.2, R3.4, R3.5
Calculate and interpret residuals.	3.2	169	R3.3, R3.4
Explain the concept of least squares.	3.2	Discussion on 169	R3.5
Determine the equation of a least-squares regression line using technology or computer output.	3.2	Technology Corner on 171, 181	R3.3, R3.4
Construct and interpret residual plots to assess whether a linear model is appropriate.	3.2	Discussion on 175, 180	R3.3, R3.4
Interpret the standard deviation of the residuals and r^2 and use these values to assess how well the least-squares regression line models the relationship between two variables.	3.2	180	R3.3, R3.5
Describe how the slope, y intercept, standard deviation of the residuals, and r^2 are influenced by outliers.	3.2	Discussion on 188	R3.1
Find the slope and y intercept of the least-squares regression line from the means and standard deviations of x and y and their correlation.	3.2	183	R3.5

Chapter 3 Chapter Review Exercises

These exercises are designed to help you review the important ideas and methods of the chapter.

R3.1 Born to be old? Is there a relationship between the gestational period (time from conception to birth) of an animal and its average life span? The figure shows a scatterplot of the gestational period and average life span for 43 species of animals.[30]



(a) Describe the association shown in the scatterplot.

(b) Point A is the hippopotamus. What effect does this point have on the correlation, the equation of the least-squares regression line, and the standard deviation of the residuals?

(c) Point B is the Asian elephant. What effect does this point have on the correlation, the equation of the least-squares regression line, and the standard deviation of the residuals?

R3.2 Penguins diving A study of king penguins looked for a relationship between how deep the penguins dive to seek food and how long they stay under water.[31] For all but the shallowest dives, there is a linear relationship that is different for different penguins. The study gives a scatterplot for one penguin titled "The Relation of Dive Duration (y) to Depth (x)." Duration y is measured in minutes and depth x is in meters. The report then says, "The regression equation for this bird is: $\hat{y} = 2.69 + 0.0138x$."

(a) What is the slope of the regression line? Interpret this value.

(b) Does the y intercept of the regression line make any sense? If so, interpret it. If not, explain why not.

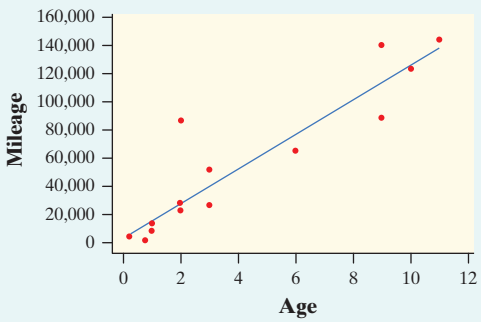
(c) According to the regression line, how long does a typical dive to a depth of 200 meters last?

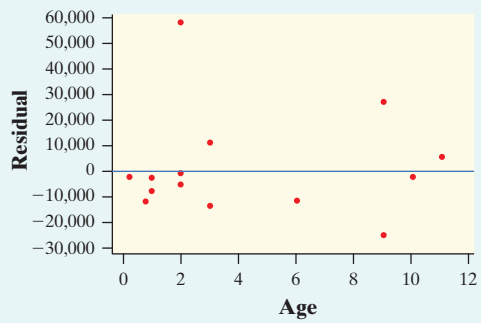
(d) Suppose that the researchers reversed the variables, using x = dive duration and y = depth. What effect will this have on the correlation? On the equation of the least-squares regression line?

R3.3 Stats teachers' cars A random sample of AP® Statistics teachers was asked to report the age (in years) and mileage of their primary vehicles. A scatterplot of the data, a least-squares regression printout, and a residual plot are provided below.

Predictor	Coef	SE Coef	T	P
Constant	3704	8268	0.45	0.662
Age	12188	1492	8.17	0.000

S = 20870.5 R-Sq = 83.7% R-Sq(adj) = 82.4%





(a) Give the equation of the least-squares regression line for these data. Identify any variables you use.

(b) One teacher reported that her 6-year-old car had 65,000 miles on it. Find and interpret its residual.

(c) What's the correlation between car age and mileage? Interpret this value in context.

(d) Is a linear model appropriate for these data? Explain how you know.

(e) Interpret the values of s and r^2.

R3.4 Late bloomers? Japanese cherry trees tend to blossom early when spring weather is warm and later when spring weather is cool. Here are some data on the average March temperature (in °C) and the day in April when the first cherry blossom appeared over a 24-year period:[32]

Temperature (°C):	4.0	5.4	3.2	2.6	4.2	4.7	4.9	4.0	4.9	3.8	4.0	5.1
Days in April to first bloom:	14	8	11	19	14	14	14	21	9	14	13	11
Temperature (°C):	4.3	1.5	3.7	3.8	4.5	4.1	6.1	6.2	5.1	5.0	4.6	4.0
Days in April to first bloom:	13	28	17	19	10	17	3	3	11	6	9	11

(a) Make a well-labeled scatterplot that's suitable for predicting when the cherry trees will bloom from the temperature. Which variable did you choose as the explanatory variable? Explain.

(b) Use technology to calculate the correlation and the equation of the least-squares regression line. Interpret the correlation, slope, and y intercept of the line in this setting.

(c) Suppose that the average March temperature this year was 8.2°C. Would you be willing to use the equation in part (b) to predict the date of first bloom? Explain.

(d) Calculate and interpret the residual for the year when the average March temperature was 4.5°C. Show your work.

(e) Use technology to help construct a residual plot. Describe what you see.

R3.5 **What's my grade?** In Professor Friedman's economics course, the correlation between the students' total scores prior to the final examination and their final-examination scores is $r = 0.6$. The pre-exam totals for all students in the course have mean 280 and standard deviation 30. The final-exam scores have mean 75 and standard deviation 8. Professor Friedman has lost Julie's final exam but knows that her total before the exam was 300. He decides to predict her final-exam score from her pre-exam total.

(a) Find the equation for the appropriate least-squares regression line for Professor Friedman's prediction.

(b) Use the least-squares regression line to predict Julie's final-exam score.

(c) Explain the meaning of the phrase "least squares" in the context of this question.

(d) Julie doesn't think this method accurately predicts how well she did on the final exam. Determine r^2. Use this result to argue that her actual score could have been much higher (or much lower) than the predicted value.

R3.6 **Calculating achievement** The principal of a high school read a study that reported a high correlation between the number of calculators owned by high school students and their math achievement. Based on this study, he decides to buy each student at his school two calculators, hoping to improve their math achievement. Explain the flaw in the principal's reasoning.

Chapter 3 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

T3.1 A school guidance counselor examines the number of extracurricular activities that students do and their grade point average. The guidance counselor says, "The evidence indicates that the correlation between the number of extracurricular activities a student participates in and his or her grade point average is close to zero." A correct interpretation of this statement would be that

(a) active students tend to be students with poor grades, and vice versa.

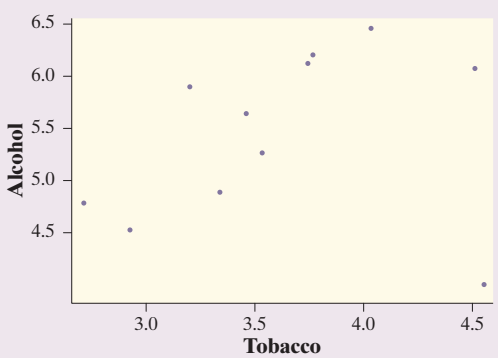
(b) students with good grades tend to be students who are not involved in many extracurricular activities, and vice versa.

(c) students involved in many extracurricular activities are just as likely to get good grades as bad grades; the same is true for students involved in few extracurricular activities.

(d) there is no linear relationship between number of activities and grade point average for students at this school.

(e) involvement in many extracurricular activities and good grades go hand in hand.

T3.2 The British government conducts regular surveys of household spending. The average weekly household spending (in pounds) on tobacco products and alcoholic beverages for each of 11 regions in Great Britain was recorded. A scatterplot of spending on alcohol versus spending on tobacco is shown below. Which of the following statements is true?



(a) The observation (4.5, 6.0) is an outlier.

(b) There is clear evidence of a negative association between spending on alcohol and tobacco.

(c) The equation of the least-squares line for this plot would be approximately $\hat{y} = 10 - 2x$.

(d) The correlation for these data is $r = 0.99$.

(e) The observation in the lower-right corner of the plot is influential for the least-squares line.

T3.3 The fraction of the variation in the values of y that is explained by the least-squares regression of y on x is

(a) the correlation.
(b) the slope of the least-squares regression line.
(c) the square of the correlation coefficient.
(d) the intercept of the least-squares regression line.
(e) the residual.

T3.4 An AP® Statistics student designs an experiment to see whether today's high school students are becoming too calculator-dependent. She prepares two quizzes, both of which contain 40 questions that are best done using paper-and-pencil methods. A random sample of 30 students participates in the experiment. Each student takes both quizzes—one with a calculator and one without—in a random order. To analyze the data, the student constructs a scatterplot that displays the number of correct answers with and without a calculator for each of the 30 students. A least-squares regression yields the equation

$$\widehat{\text{Calculator}} = -1.2 + 0.865(\text{Pencil}) \quad r = 0.79$$

Which of the following statements is/are true?

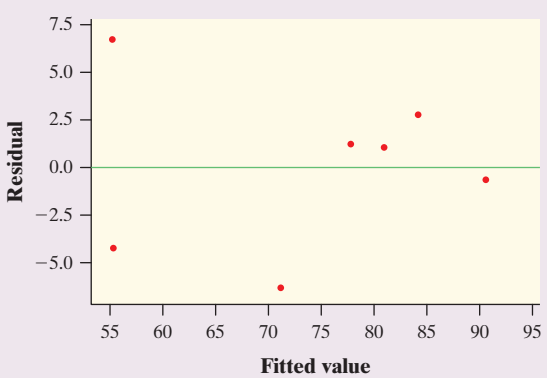
I. If the student had used Calculator as the explanatory variable, the correlation would remain the same.
II. If the student had used Calculator as the explanatory variable, the slope of the least-squares line would remain the same.
III. The standard deviation of the number of correct answers on the paper-and-pencil quizzes was larger than the standard deviation on the calculator quizzes.

(a) I only (c) III only (e) I, II, and III
(b) II only (d) I and III only

Questions T3.5 and T3.6 refer to the following setting. Scientists examined the activity level of 7 fish at different temperatures. Fish activity was rated on a scale of 0 (no activity) to 100 (maximal activity). The temperature was measured in degrees Celsius. A computer regression printout and a residual plot are given below. Notice that the horizontal axis on the residual plot is labeled "Fitted value."

Predictor	Coef	SE Coef	T	P
Constant	148.62	10.71	13.88	0.000
Temperature	-3.2167	0.4533	-7.10	0.001

S = 4.78505 R-Sq = 91.0% R-Sq(adj) = 89.2%



T3.5 What was the activity level rating for the fish at a temperature of 20°C?

(a) 87 (b) 84 (c) 81 (d) 66 (e) 3

T3.6 Which of the following gives a correct interpretation of s in this setting?

(a) For every 1°C increase in temperature, fish activity is predicted to increase by 4.785 units.
(b) The typical distance of the temperature readings from their mean is about 4.785°C.
(c) The typical distance of the activity level ratings from the least-squares line is about 4.785 units.
(d) The typical distance of the activity level readings from their mean is about 4.785.
(e) At a temperature of 0°C, this model predicts an activity level of 4.785.

T3.7 Which of the following statements is *not* true of the correlation r between the lengths in inches and weights in pounds of a sample of brook trout?

(a) r must take a value between -1 and 1.
(b) r is measured in inches.
(c) If longer trout tend to also be heavier, then $r > 0$.
(d) r would not change if we measured the lengths of the trout in centimeters instead of inches.
(e) r would not change if we measured the weights of the trout in kilograms instead of pounds.

T3.8 When we standardize the values of a variable, the distribution of standardized values has mean 0 and standard deviation 1. Suppose we measure two variables X and Y on each of several subjects. We standardize both variables and then compute the least-squares regression line. Suppose the slope of the least-squares regression line is -0.44. We may conclude that

(a) the intercept will also be -0.44.
(b) the intercept will be 1.0.
(c) the correlation will be $1/-0.44$.
(d) the correlation will be 1.0.
(e) the correlation will also be -0.44.

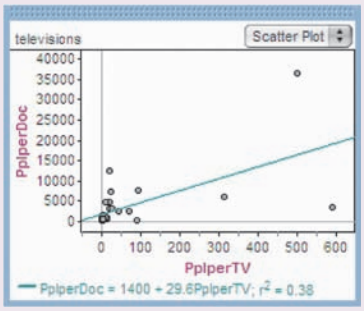
T3.9 There is a linear relationship between the number of chirps made by the striped ground cricket and the air temperature. A least-squares fit of some data collected by a biologist gives the model $\hat{y} = 25.2 + 3.3x$, where x is the number of chirps per minute and $\hat{y}$ is the estimated temperature in degrees Fahrenheit. What is the predicted increase in temperature for an increase of 5 chirps per minute?

(a) 3.3°F (c) 25.2°F (e) 41.7°F
(b) 16.5°F (d) 28.5°F

T3.10 A data set included the number of people per television set and the number of people per physician for 40 countries. The Fathom screen shot below displays

a scatterplot of the data with the least-squares regression line added. In Ethiopia, there were 503 people per TV and 36,660 people per doctor. What effect would removing this point have on the regression line?

(a) Slope would increase; y intercept would increase.
(b) Slope would increase; y intercept would decrease.
(c) Slope would decrease; y intercept would increase.
(d) Slope would decrease; y intercept would decrease.
(e) Slope and y intercept would stay the same.



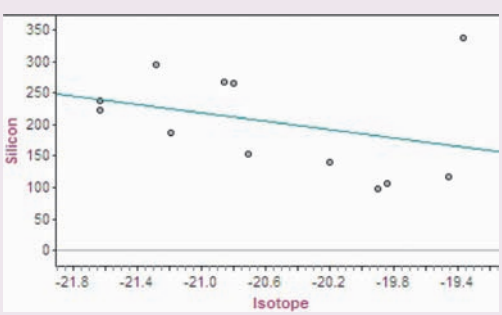
Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

T3.11 Sarah's parents are concerned that she seems short for her age. Their doctor has the following record of Sarah's height:

Age (months):	36	48	51	54	57	60
Height (cm):	86	90	91	93	94	95

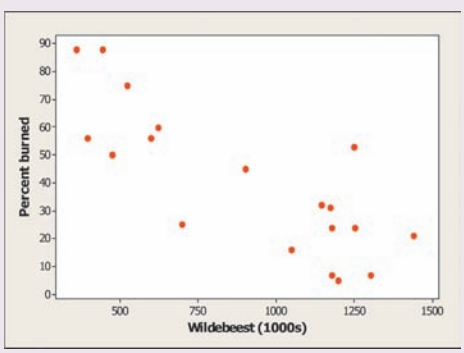
(a) Make a scatterplot of these data.
(b) Using your calculator, find the equation of the least-squares regression line of height on age.
(c) Use your regression line to predict Sarah's height at age 40 years (480 months). Convert your prediction to inches (2.54 cm = 1 inch).
(d) The prediction is impossibly large. Explain why this happened.

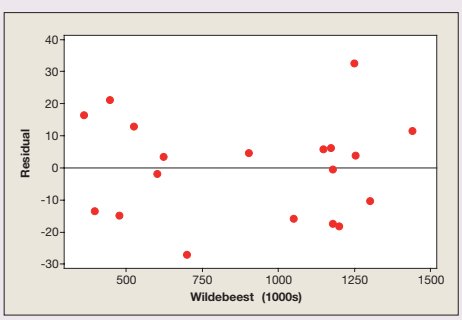
T3.12 Drilling down beneath a lake in Alaska yields chemical evidence of past changes in climate. Biological silicon, left by the skeletons of single-celled creatures called diatoms, is a measure of the abundance of life in the lake. A rather complex variable based on the ratio of certain isotopes relative to ocean water gives an indirect measure of moisture, mostly from snow. As we drill down, we look further into the past. Here is a scatterplot of data from 2300 to 12,000 years ago:



(a) Identify the unusual point in the scatterplot. Explain what's unusual about this point.
(b) If this point was removed, describe the effect on
  i. the correlation.
  ii. the slope and y intercept of the least-squares line.
  iii. the standard deviation of the residuals.

T3.13 Long-term records from the Serengeti National Park in Tanzania show interesting ecological relationships. When wildebeest are more abundant, they graze the grass more heavily, so there are fewer fires and more trees grow. Lions feed more successfully when there are more trees, so the lion population increases. Researchers collected data on one part of this cycle, wildebeest abundance (in thousands of animals) and the percent of the grass area burned in the same year. The results of a least-squares regression on the data are shown here.[33]



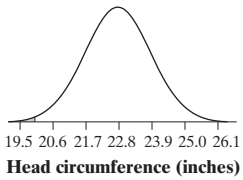


Predictor	Coef	SE Coef	T	P
Constant	92.29	10.06	9.17	0.000
Wildebeest (1000s)	−0.05762	0.01035	−5.56	0.000

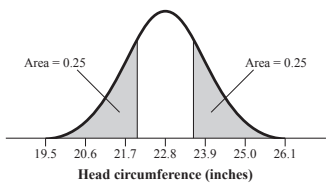
S = 15.9880 R-Sq = 64.6% R-Sq(adj) = 62.5%

(a) Give the equation of the least-squares regression line. Be sure to define any variables you use.
(b) Explain what the slope of the regression line means in this setting.
(c) Find the correlation. Interpret this value in context.
(d) Is a linear model appropriate for describing the relationship between wildebeest abundance and percent of grass area burned? Support your answer with appropriate evidence.

circumferences less than 20 inches or greater than 26 inches and require custom helmets.



(c) For male soldiers, head circumference follows a $N(22.8, 1.1)$ distribution. The 1st quartile is the boundary value with 25% of the area to its left. The 3rd quartile is the boundary value with 75% of the area to its left (see graph below). A z-score of -0.67 gives the value closest to 0.25 (0.2514). Solving $-0.67 = \dfrac{x - 22.8}{1.1}$ gives $Q_1 = 22.063$. A z-score of 0.67 gives the value closest to 0.75 (0.7486). Solving $0.67 = \dfrac{x - 22.8}{1.1}$ gives $Q_3 = 23.537$. *Using technology:* `invNorm(area:0.25,μ:22.8,σ:1.1)` gives $Q_1 = 22.058$ and `invNorm(area:0.75,μ:22.8,σ:1.1)` gives $Q_3 = 23.542$. Thus, $IQR = 23.542 - 22.058 = 1.484$ inches.



T2.13 No. First, there is a large difference between the mean and the median. In a Normal distribution, the mean and median are the same, but in this distribution the mean is 48.25 and the median is 37.80. Second, the distance between the minimum and the median is 35.80 but the distance between the median and the maximum is 167.10. In a Normal distribution, these distances should be about the same. Both of these facts suggest that the distribution is skewed to the right.

Chapter 3

Section 3.1

Answers to Check Your Understanding

page 144: **1.** Explanatory: number of cans of beer. Response: blood alcohol level. **2.** Explanatory: amount of debt and income. Response: stress caused by college debt.

page 149: **1.** Positive. The longer the duration of the eruption, the longer we should expect to wait between eruptions because long eruptions use more energy and it will take longer to build up the energy needed to erupt again. **2.** Roughly linear with two clusters. The clusters indicate that, in general, there are two types of eruptions—shorter eruptions that last around 2 minutes and longer eruptions that last around 4.5 minutes. **3.** Fairly strong. The points don't deviate much from the linear form. **4.** There are a few possible outliers around the clusters. However, there aren't many and potential outliers are not very distant from the main clusters of points. **5.** How long the previous eruption was.

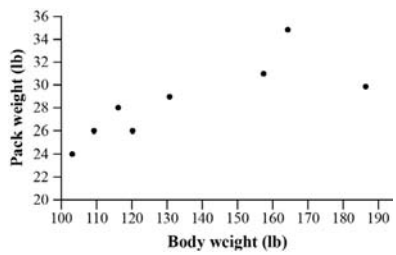
page 153: **(a)** $r \approx 0.9$. This indicates that there is a strong, positive linear relationship between the number of boats registered in Florida and the number of manatees killed. **(b)** $r \approx 0.5$. This indicates that there is a moderate, positive linear relationship between the number of named storms predicted and the actual number of named storms. **(c)** $r \approx 0.3$. This indicates that there is a weak, positive linear relationship between the healing rate of the two front limbs of the newts. **(d)** $r \approx -0.1$. This indicates that there is a weak, negative linear relationship between last year's percent return and this year's percent return in the stock market.

Answers to Odd-Numbered Section 3.1 Exercises

3.1 Explanatory: water temperature (quantitative). Response: weight change (quantitative).

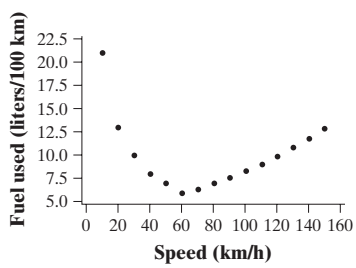
3.3 **(a)** Positive. Students with higher IQs tend to have higher GPAs and vice versa because both IQ and GPA are related to mental ability. **(b)** Roughly linear, because a line through the scatterplot of points would provide a good summary. Moderately strong, because most of the points would be close to the line. **(c)** IQ $\approx$ 103 and GPA $\approx$ 0.4.

3.5 A scatterplot is shown below.



3.7 **(a)** There is a positive association between backpack weight and body weight. For students under 140 pounds, there seems to be a linear pattern in the graph. However, for students above 140 pounds, the association begins to curve. Because the points vary somewhat from the linear pattern, the relationship is only moderately strong. **(b)** The hiker with body weight 187 pounds and pack weight 30 pounds. This hiker makes the form appear to be nonlinear for weights above 140 pounds. Without this hiker, the association would look very linear for all body weights.

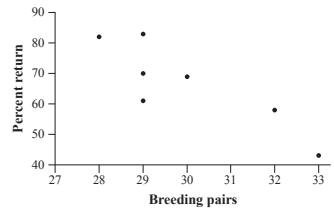
3.9 **(a)** A scatterplot is shown below. **(b)** The relationship is curved. Large amounts of fuel were used for low and high values of speed and smaller amounts of fuel were used for moderate speeds. This makes sense because the best fuel efficiency is obtained by driving at moderate speeds. **(c)** Both directions are present in the scatterplot. The association is negative for lower speeds and positive for higher speeds. **(d)** The relationship is very strong, with little deviation from a curve that can be drawn through the points.



3.11 **(a)** Most of the southern states fall in the same pattern as the rest of the states. However, southern states typically have lower mean SAT math scores than other states with a similar percent of students taking the SAT. **(b)** West Virginia has a much lower mean

SAT Math score than the other states that have a similar percent of students taking the exam.

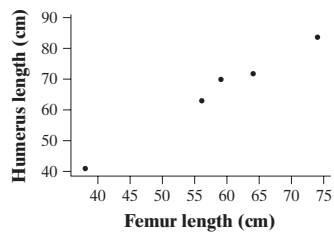
3.13 A scatterplot is shown below. There is a negative, linear, moderately strong relationship between the percent returning and the number of breeding pairs.



3.15 (a) $r = 0.9$ (b) $r = 0$ (c) $r = 0.7$ (d) $r = -0.3$ (e) $r = -0.9$

3.17 (a) Gender is a categorical variable and correlation r is for two quantitative variables. (b) The largest possible value of the correlation is $r = 1$. (c) The correlation r has no units.

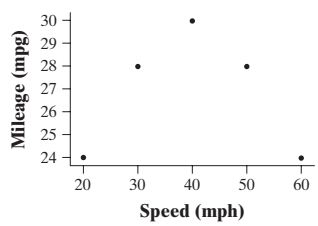
3.19 (a) The scatterplot below shows a strong, positive linear relationship between the two measurements. It appears that all five specimens come from the same species. (b) The femur measurements have $\bar{x} = 58.2$ and $s_x = 13.2$. The humerus measurements have $\bar{y} = 66$ and $s_y = 15.89$. The sum of the z-score products is 3.97620, so the correlation coefficient is $r = (1/4)(3.97620) = 0.9941$. The very high value of the correlation confirms the strong, positive linear association between femur length and humerus length in the scatterplot from part (a).



3.21 (a) There is a strong, positive linear association between sodium and calories. (b) It increases the correlation. It falls in the linear pattern of the rest of the data and observations with unusually small or unusually large values of x have a big influence on the correlation.

3.23 (a) The correlation would not change, because correlation is not affected by a change of units for either variable. (b) The correlation would not change, because it does not distinguish between explanatory and response variables.

3.25 (a) A scatterplot is shown below. (b) $r = 0$ (c) The correlation measures the strength of a *linear* association, but this plot shows a nonlinear relationship between speed and mileage.

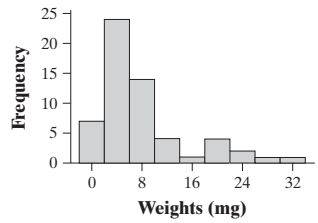


3.27 a

3.29 d

3.31 b

3.33 A histogram is shown below. The distribution is right-skewed, with several possible high outliers. Because of the skewness and outliers, we should use the median (5.4 mg) and *IQR* (5.5 mg) to describe the center and spread.



Section 3.2

Answers to Check Your Understanding

page 168: **1.** 40. For each additional week, we predict that a rat will gain 40 grams of weight. **2.** 100. The predicted weight for a newborn rat is 100 grams. **3.** $\hat{y} = 100 + 40(16) = 740$ grams **4.** 2 years = 104 weeks, so $\hat{y} = 100 + 40(104) = 4260$ grams. This is equivalent to 9.4 pounds (about the weight of a large newborn human). This is unreasonable and is the result of extrapolation.

page 172: The answer is given in the text.

page 174: **1.** $y - \hat{y} = 31{,}891 - 36{,}895 = -\$5004$ **2.** The actual price of this truck is \$5004 less than predicted based on the number of miles it has been driven. **3.** The truck with 44,447 miles and a price of \$22,896. This truck has a residual of −\$8120, which means that the line overpredicted the price by \$8120. No other truck had a residual that was farther below 0 than this one.

page 176: **1.** The backpack for this hiker was almost 4 pounds heavier than expected based on the weight of the hiker. **2.** Because there appears to be a negative-positive-negative pattern in the residual plot, a linear model is not appropriate for these data.

Answers to Odd-Numbered Section 3.2 Exercises

3.35 predicted weight = 80 − 6 (days)

3.37 (a) 1.109. For each 1-mpg increase in city mileage, the predicted highway mileage will increase by 1.109 mpg. (b) 4.62 mpg. This would represent the highway mileage for a car that gets 0 mpg in the city, which is impossible. (c) 22.36 mpg

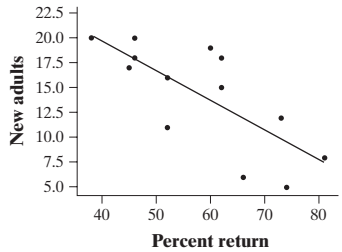
3.39 (a) −0.0053. For each additional week in the study, the predicted pH decreased by 0.0053 units. (b) 5.43. The predicted pH level at the beginning of the study (weeks = 0) is 5.43. (c) 4.635

3.41 No. 1000 months is well outside the observed time period and we can't be sure that the linear relationship continues after 150 weeks.

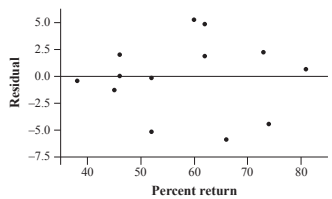
3.43 The line $\hat{y} = 1 - x$ is a much better fit. The sum of squared residuals for this line is only 3, while the sum of squared residuals for $\hat{y} = 3 - 2x$ is 18.

3.45 residual = 5.08 − 5.165 = −0.085. The actual pH value for that week was 0.085 less than predicted.

3.47 (a) The scatterplot (with regression line) is shown below. (b) $\hat{y} = 31.9 - 0.304x$. (c) For each increase of 1 in the percent of returning birds, the predicted number of new adult birds will decrease by 0.304. (d) residual = 11 − 16.092 = −5.092. In this colony, there were 5.092 fewer new adults than expected based on the percent of returning birds.



3.49 **(a)** Because there is no obvious leftover pattern in the residual plot shown below, a line is an appropriate model to use for these data. **(b)** The point with the largest residual (66% returning) has a residual of about −6. This means that the colony with 66% returning birds has about 6 fewer new adults than predicted based on the percent returning.



3.51 No. Because there is an obvious negative-positive-negative pattern in the residual plot, a linear model is not appropriate for these data.

3.53 **(a)** There is a positive, linear association between the two variables. There is more variation in the field measurements for larger laboratory measurements. **(b)** No. The points for the larger depths fall systematically below the line $y = x$, showing that the field measurements are too small compared to the laboratory measurements. **(c)** The slope would be closer to 0 and the y intercept would be larger.

3.55 **(a)** residual = 150.06 − 146.295 = 3.765. Yu-Na Kim's free skate score was 3.765 points higher than predicted based on her short program score. **(b)** Because there is no leftover pattern in the residual plot, a linear model is appropriate for these data. **(c)** When using the least-squares regression line with x = short program score to predict y = free skate score, we will typically be off by about 10.2 points. **(d)** About 73.6% of the variation in free skate scores is accounted for by the linear model relating free skate scores to short program scores.

3.57 r^2: About 56% of the variation in the number of new adults is accounted for by the linear model relating number of new adults to the percent returning. s: When using the least-squares regression line with x = percent returning to predict y = number of new adults, we will typically be off by 3.67 adults.

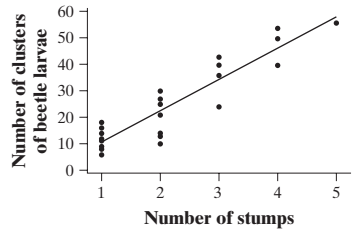
3.59 **(a)** $\hat{y} = 266.07 - 6.650x$, where y = percent of males that return the next year and x = number of breeding pairs. When $x = 30$, $\hat{y} = 66.57$. **(b)** R-Sq = 74.6% **(c)** $r = -\sqrt{0.746} = -0.864$. The sign is negative because the slope is negative. **(d)** When using the least-squares regression line with x = number of breeding pairs to predict y = percent returning, we will typically be off by 7.76%.

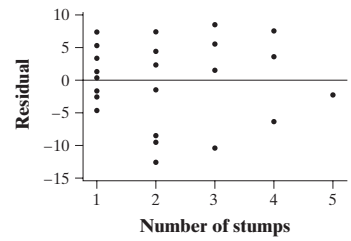
3.61 **(a)** $\hat{y} = 33.67 + 0.54x$. **(b)** If the value of x is 1 standard deviation below $\bar{x}$, the predicted value of y will be r standard deviations of y below $\bar{y}$. So the predicted value for the husband is 68.5 − 0.5(2.7) = 67.15 inches.

3.63 **(a)** $r^2 = 0.25$. About 25% of the variation in husbands' heights is accounted for by the linear model relating husband's height to wife's height. **(b)** When using the least-squares regression line with x = wife's height to predict y = husband's height, we will typically be off by 1.2 inches.

3.65 **(a)** $\hat{y} = x$ where y = final and x = midterm **(b)** If $x = 50$, $\hat{y} = 67.1$. If $x = 100$, $\hat{y} = 87.6$. **(c)** The student who did poorly on the midterm (50) is predicted to do better on the final (closer to the mean), while the student who did very well on the midterm (100) is predicted to do worse on the final (closer to the mean).

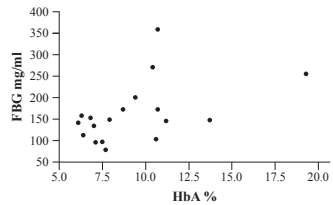
3.67 *State*: Is a linear model appropriate for these data? If so, how well does the least-squares regression line fit the data? *Plan*: We will look at the scatterplot and residual plot to see if the association is linear or nonlinear. Then, if a linear model is appropriate, we will use s and r^2 to measure how well the line fits the data. *Do*: The scatterplot below shows a moderately strong, positive linear association between the number of stumps and the number of clusters of beetle larvae. The residual plot doesn't show any obvious leftover pattern, confirming that a linear model is appropriate.





$\hat{y} = -1.29 + 11.89x$, where y = number of clusters of beetle larvae and x = number of stumps. $s = 6.42$, meaning that our predictions will typically be off by about 6.42 clusters when we use the line to predict the number of clusters of beetle larvae from the number of stumps. Finally, $r^2 = 0.839$, meaning 83.9% of the variation in the number of clusters of beetle larvae is accounted for by the linear model relating number of clusters of beetle larvae to the number of stumps. *Conclude*: The linear model relating number of clusters of beetle larvae to the number of stumps is appropriate and fits the data well, accounting for more than 80% of the variation in number of clusters of beetle larvae.

3.69 **(a)** A scatterplot is shown below. There is a moderate, positive linear association between HbA and FBG. There are possible outliers to the far right (subject 18) and near the top of the plot (subject 15).



(b) Because the point is in the positive, linear pattern formed by most of the data values, it makes r closer to 1. Also, because the point is likely to be below the least-squares regression line, it will "pull down" the line on the right side, making the slope closer to 0. Without the outlier, r decreases from 0.4819 to 0.3837 as expected. Likewise, the equation changes from $\hat{y} = 66.4 + 10.4x$ to $\hat{y} = 52.3 + 12.1x$. **(c)** The point makes r closer to 0 because it is out of the linear pattern formed by most of the data values. Because this point's x coordinate is very close to $\bar{x}$ but the y coordinate is far above $\bar{y}$, it won't influence the slope very much but will increase the y intercept. Without the outlier, r increases from 0.4819 to 0.5684, as expected. Likewise, the equation changes from $\hat{y} = 66.4 + 10.4x$ to $\hat{y} = 69.5 + 8.92x$.

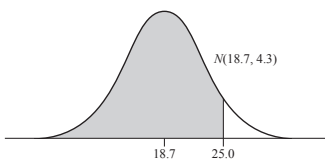
3.71 a

3.73 c

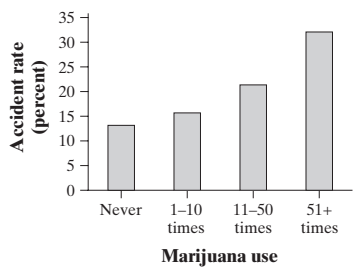
3.75 d

3.77 b

3.79 For these vehicles, the combined mileage follows a N(18.7, 4.3) distribution and we want to find the percent of cars with lower mileage than 25 (see graph below). $z = \dfrac{25 - 18.7}{4.3} = 1.47$. From Table A, the proportion of z-scores below 1.47 is 0.9292. *Using technology*: `normalcdf(lower:-1000,upper:25,μ:18.7,σ:4.3)` = 0.9286. About 93% percent of vehicles get worse combined mileage than the Chevrolet Malibu.



3.81 **(a)** A bar graph is given below. The people who use marijuana more are more likely to have caused accidents. **(b)** Association does not imply causation. For example, it could be that drivers who use marijuana more often are more willing to take risks than other drivers and that the willingness to take risks is what is causing the higher accident rate.



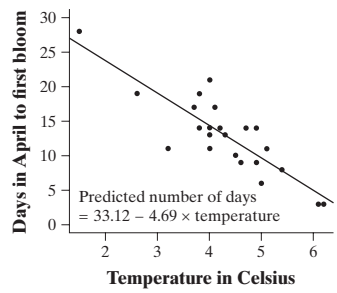
Answers to Chapter 3 Review Exercises

R3.1 **(a)** There is a moderate, positive linear association between gestation and life span. Without the outliers at the top and in the upper right, the association appears moderately strong, positive, and curved. **(b)** It makes r closer to 0 because it decreases the strength of what would otherwise be a moderately strong positive association. Because this point is close to $\bar{x}$ but far above $\bar{y}$, it won't affect the slope much but will increase the y intercept. Because it has such a large residual, it increases s. **(c)** Because it is in the positive, linear pattern formed by most of the data values, it will make r closer to 1. Also, because the point is likely to be above the least-squares regression line, it will "pull up" the line on the right side, making the slope larger and the intercept smaller. Because this point is likely to have a small residual, it decreases s.

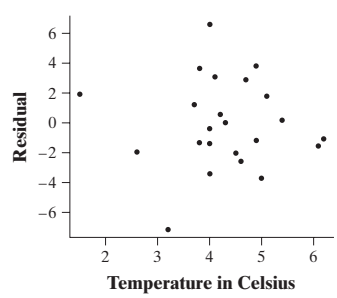
R3.2 **(a)** 0.0138. For each increase of 1 meter in dive depth, the predicted duration increases by 0.0138 minutes. **(b)** The y intercept suggests that a dive of 0 depth would last an average of 2.69 minutes; this obviously does not make any sense. **(c)** 5.45 minutes **(d)** If the variables are reversed, the correlation will remain the same. However, the slope and y intercept will be different.

R3.3 **(a)** $\hat{y} = 3704 + 12{,}188x$, where y represents the mileage of the cars and x represents the age. **(b)** residual = 65,000 − 76,832 = −11,832. This teacher has driven 11,832 fewer miles than predicted based on the age of the car. **(c)** $r = +\sqrt{0.837} = 0.915$. This shows that there is a strong, positive linear association between the age of cars and their mileage. **(d)** Yes, because there is no leftover pattern in the residual plot. **(e)** s = 20,870.5: When using the least-squares regression line with x = car's age to predict y = number of miles it has been driven, we will typically be off by about 20,870.5 miles. r^2 = 83.7%: About 83.7% of the variability in mileage is accounted for by the linear model relating mileage to age.

R3.4 **(a)** The scatterplot is shown below. Average March temperature, because changes in March temperature probably have an effect on the date of first bloom.



(b) $r = -0.85$ and $\hat{y} = 33.12 - 4.69x$, where y represents the number of days and x represents the temperature. r: There is a strong, negative linear association between the average March temperature and the days in April until first bloom. *Slope*: For every 1° increase in average March temperature, the predicted number of days in April until first bloom decreases by 4.69. *y intercept*: If the average March temperature was 0°C, the predicted number of days in April to first bloom is 33.12 (May 3). **(c)** No, x = 8.2 is well beyond the values of x we have in the data set. **(d)** residual = 10 − 12.015 = −2.015. In this year, the actual date of first bloom occurred about 2 days earlier than predicted based on the average March temperature. **(e)** There is no leftover pattern in the residual plot shown below, indicating that a linear model is appropriate.



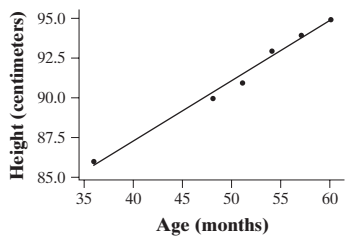
R3.5 (a) $\hat{y} = 30.2 + 0.16x$, where y = final exam score and x = total score before the final examination. (b) 78.2 (c) Of all the lines that the professor could use to summarize the relationship between final exam score and total points before the final exam, the least-squares regression line is the one that has the smallest sum of squared residuals. (d) Because $r^2 = 0.36$, only 36% of the variability in the final exam scores is accounted for by the linear model relating final exam scores to total score before the final exam. More than half (64%) of the variation in final exam scores is *not* accounted for, so Julie has reason to question this estimate.

R3.6 Even though there is a high correlation between number of calculators and math achievement, we shouldn't conclude that increasing the number of calculators will *cause* an increase in math achievement. It is possible that students who are more serious about school have better math achievement and also have more calculators.

Answers to Chapter 3 AP® Statistics Practice Test

T3.1 d
T3.2 e
T3.3 c
T3.4 a
T3.5 a
T3.6 c
T3.7 b
T3.8 e
T3.9 b
T3.10 c

T3.11 (a) A scatterplot with regression line is shown below. (b) $\hat{y} = 71.95 + 0.3833x$, where y = height and x = age. (c) 255.934 cm, or 100.76 inches (d) This was an extrapolation. Our data were based only on the first 5 years of life and the linear trend will not continue forever.



T3.12 (a) The point in the upper-right-hand corner has a very high silicon value for its isotope value. (b) (i) r would get closer to -1 because it does not follow the linear pattern of the other points. (ii) Because this point is "pulling up" the line on the right side of the plot, removing it will make the slope steeper (more negative) and the y intercept smaller (note that the y axis is to the *right* of the points in the scatterplot). (iii) Because this point has a large residual, removing it will make s a little smaller.

T3.13 (a) $\hat{y} = 92.29 - 0.05762x$, where y is the percent of the grass burned and x is the number of wildebeest. (b) For every increase of 1000 wildebeest, the predicted percent of grassy area burned decreases by about 0.058. (c) $r = -\sqrt{0.646} = -0.804$. There is a strong, negative linear association between the percent of grass burned and the number of wildebeest. (d) Yes, because there is no obvious leftover pattern in the residual plot.

Chapter 4

Section 4.1

Answers to Check Your Understanding

page 213: **1.** Convenience sampling. This could lead the inspector to overestimate the quality of the oranges if the farmer puts the best oranges on top. **2.** Voluntary response sampling. In this case, those who are happy that the UN has its headquarters in the U.S. already have what they want and so are less likely to respond. The proportion who answered "No" in the sample is likely to be higher than the true proportion in the U.S. who would answer "No."

page 223: **1.** You would have to identify 200 different seats, go to those seats in the arena, and find the people who are sitting there, which would take a lot of time. **2.** It is best to create strata where the people within a stratum are very similar to each other but different than the people in other strata. In this case, it would be better to take the lettered rows as the strata because each lettered row is the same distance from the court and so would contain only seats with the same (or nearly the same) ticket price. **3.** It is best if the people in each cluster reflect the variability found in the population. In this case, it would be better to take the numbered sections as the clusters because they include all different seat prices.

page 228: **1.** (a) Undercoverage (b) Nonresponse (c) Undercoverage **2.** By making it sound like they are not a problem in the landfill, this question will result in fewer people suggesting that we should ban disposable diapers. The proportion who would say "Yes" to this survey question is likely to be smaller than the proportion who would say "Yes" to a more fairly worded question.

Answers to Odd-Numbered Section 4.1 Exercises

4.1 Population: all local businesses. Sample: the 73 businesses that return the questionnaire.

4.3 Population: the 1000 envelopes stuffed during a given hour. Sample: the 40 randomly selected envelopes.

4.5 This is a voluntary response sample. In this case, it appears that people who strongly support gun control volunteered more often, causing the proportion in the sample to be greater than the proportion in the population.

4.7 This is a voluntary response sample and overrepresents the opinions of those who feel most strongly about the issue being surveyed.

4.9 (a) A convenience sample (b) The first 100 students to arrive at school likely had to wake up earlier than other students, so 7.2 hours is probably less than the true average.

4.11 (a) Number the 40 students from 01 to 40. Pick a starting point on the random number table. Record two-digit numbers, skipping numbers that aren't between 01 and 40 and any repeated numbers, until you have 5 unique numbers between 01 and 40. Use the 5 students corresponding to these numbers. (b) Using line 107, skip the numbers not in bold: 82 73 95 78 90 **20** 80 74 75 **11** 81 67 65 53 00 94 **38** **31** 48 93 60 94 **07**. Select Johnson (20), Drasin (11), Washburn (38), Rider (31), and Calloway (07).

4.13 (a) *Using calculator*: Number the plots from 1 to 1410. Use the command `randInt(1,1410)` to select 141 different integers from 1 to 1410 and use the corresponding 141 plots. (b) Answers will vary.

4.15 (a) False—although, on average, there will be four 0s in every set of 40 digits, the number of 0s can be less than 4 or greater than 4 by chance. (b) True—there are 100 pairs of digits 00 through 99,